



TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

Antti Liski

**Statistical and Information Theoretic Approaches to
Model Selection and Averaging**



Julkaisu 1120 • Publication 1120

Tampere 2013

Antti Liski

Statistical and Information Theoretic Approaches to Model Selection and Averaging

Thesis for the degree of Doctor of Philosophy to be presented with due permission for public examination and criticism in Tietotalo Building, Auditorium TB109, at Tampere University of Technology, on the 12th of April 2013, at 12 noon.

Supervisor:

Prof. Ioan Tăbuș (Custos),
Department of Signal Processing,
Tampere University of Technology,
Tampere, Finland.

Pre-examiners:

Prof. Juha Karvanen,
Department of Mathematics and Statistics,
University of Jyväskylä,
Jyväskylä, Finland.

Prof. Seppo Pynnönen,
Department of Mathematics and Statistics,
University of Vaasa,
Vaasa, Finland.

Opponent:

Prof. Jianxin Pan,
School of Mathematics,
University of Manchester,
Manchester, United Kingdom.

ISBN 978-952-15-3041-8 (printed)
ISBN 978-952-15-3054-8 (PDF)
ISSN 1459-2045

Abstract

In this thesis we consider model selection (MS) and its alternative, model averaging (MA), in seven research articles and in an introductory summary of the articles. The utilization of the minimum description length (MDL) principle is a common theme in five articles. In three articles we approach MA by estimating model weights using MDL and by making use of the idea of shrinkage estimation with special emphasis on the weighted average least squares (WALS) and penalized least squares (PenLS) estimation. We also apply MS and MA techniques to data on hip fracture treatment costs in seven hospital districts in Finland.

Implementation of the MDL principle for MS is put into action by using the normalized maximum likelihood (NML). However, the straightforward use of the NML technique in Gaussian linear regression fails because the normalization coefficient is not finite. Rissanen has proposed an elegant solution to the problem by constraining the data space properly. We demonstrate the effect of data constraints on the MS criterion and present a general convex constraint in data space and discuss two particular cases: the rhomboidal and ellipsoidal constraints. From these findings we derive four new NML-based criteria. One particular constraint is related to the case when collinearity is present in data.

We study WALS estimation which has the potential for a good risk profile. WALS is attractive in regression especially when the number of explanatory variables is large because its computational burden is light. We also apply WALS to estimation and comparison of hip fracture treatment costs between hospital districts in Finland. We present the WALS estimators as a special case of shrinkage estimators and we characterize a class of shrinkage estimators for which we derive the efficiency bound. We demonstrate how shrinkage estimators are obtained by using the PenLS technique and we prove sufficient conditions for the PenLS estimator to belong to the class of shrinkage estimators. Through this connection we may derive new MA estimators and effectively utilize certain previously known estimators in MA. We also study the performance of the estimators by using simulation

experiments based on hip fracture treatment cost data.

We propose an MA estimator with weights selected by the NML criterion. The resulting mixture estimator usually performs better than the corresponding MS estimator. We report on simulation experiments where the performance potential of MDL weight selection is compared with the corresponding potential of the AIC, BIC and Mallow's MA estimators. We also exploit the finding that a smoothing spline estimator may be rewritten as a linear mixed model (LMM). We present the NML criterion for LMM's and propose an automatic data-based smoothing method based on this criterion. The performance of the MDL criterion is compared to AIC, BIC and generalized cross-validation criteria in simulation experiments.

Finally we consider the sequential NML (sNML) criterion in logistic regression. We show that while the NML criterion becomes quickly computationally infeasible as the number of covariates and amount of data increases, the sNML criterion can still be exploited in MS. We also develop a risk adjustment model for hip fracture mortality in Finland by choosing comorbidities that have an effect on mortality after hip fracture.

Key words: minimum description length principle, regression, weighted average least squares, penalized least squares, shrinkage estimation, spline smoothing, hip fracture treatment costs and mortality

Acknowledgements

I wish to express my sincere gratitude to my supervisor Professor Ioan Tabus for his high level guidance and support. He has always been available and made it easy to come over and knock on his door when having problems or questions.

I also warmly thank my father Professor Emeritus Erkki Liski. There are simply too many things to thank you for to fit in these lines. You have been a supervisor, collaborator, colleague, friend and a father. Your contribution has been irreplaceable. You have come up with research problems, offered help and guidance when I have been in need and provided support and motivation throughout my life. On top of this, working with you is fun and I am privileged to learn from your vast experience. You have set the bar high both as father and as a statistician.

I thank my co-authors and especially Dr. Ciprian Doru Giurcaneanu without whom I probably would not have ended to the Tampere University of Technology (TUT) to work on this Thesis. Dr. Reijo Sund has always kindly and generously provided his help, which has been invaluable from the day we met. I also thank Professor Unto Häkkinen for continuing our collaboration after I moved from Stakes to TUT.

I wish to thank Professor Emeritus Jorma Rissanen for helpful discussions and inspiration. I also wish to thank Professor Hannu Oja for his support and for welcoming me in many of his research group's events. I thank Dr. Simo Puntanen for leading me towards an academic career first through mineral water and later through his inspiring teaching. Simo's help, enthusiasm and motivating attitude have definitely had a positive influence on my work and studies.

I am grateful to Dr. Päivi Santalahti for her supportive attitude and for giving me time to finalize this Thesis beside my work in the Yhteispeli project.

I am thankful to the pre-examiners Professor Seppo Pynnönen and Professor Juha Karvanen whose helpful comments made me rethink the summary part of this Thesis.

Finally I wish to thank my lovely wife Jutta for making every day beautiful and our wonderful sons Aleksi, Akusti, Vertti and Venni who make sure we never run out of fuss and speed. You are my never ending fountain of happiness. Last but not least I thank my mother Leena for her love and support throughout my whole life and Eero and Anni for making sure I don't get too serious.

Contents

Abstract	i
Acknowledgements	iii
Contents	v
List of original articles	vii
List of abbreviations	ix
1 Introduction	1
1.1 Approaches	1
1.1.1 Information criteria	1
1.1.2 Minimum description length principle	2
1.1.3 Regression modeling	3
1.2 Selection or averaging?	5
1.2.1 The traditional t-test	5
1.2.2 Weighted average	7
1.2.3 Least squares model averaging	8
2 Estimation and averaging	11
2.1 Least squares estimation	11
2.1.1 Restricted least squares	12
2.1.2 Averaging across restricted LS estimators	13
2.1.3 Shrinking LS estimates	14
2.1.4 Penalized least squares	17
2.2 Maximum likelihood estimation	18
2.2.1 MLE in linear mixed model	18
2.2.2 MLE in logistic regression	19

3	Model selection with MDL	21
3.1	Introduction to modeling with MDL	21
3.1.1	Prefix Codes	21
3.1.2	Codelengths and probabilities	22
3.1.3	Normalized maximum likelihood	22
3.2	Variable selection in linear regression	23
3.3	Spline smoothing	25
3.4	Sequential NML in logistic regression	25
3.5	Weight selection in nonparametric regression	27
4	Applications to hip fracture data	29
4.1	The PERFECT project	29
4.2	Medical care costs of hip fracture treatments	30
4.3	Risk-adjustment model for hip fracture mortality	30
5	Summary of publications and author's contribution	35
5.1	Summary of publications	35
5.2	Author's contribution to the articles	37
6	Conclusions	39
	References	43
	Original articles	49

List of original articles

This thesis consists of the following publications. In the text, the publications are referred to as [1],..., [7]. The publications are reproduced with kind permissions from the copyright holders.

1. Liski, E. P. and Liski, A. (2009). Minimum description length model selection in Gaussian regression under data constraints. In: Schipp, B., Krämer, W. (eds.) *Statistical Inference, Econometric Analysis and Matrix Algebra, Festschrift in Honour of Götz Trenkler*. Springer, pp.201–208.
2. Giurcaneanu, C. D., Razavi, S. A. and Liski, A. (2011). Variable selection in linear regression: Several approaches based on normalized maximum likelihood, *Signal Processing*, 91(8), pp.1671–1692.
3. Liski, A., Liski, E. P., Sund, R. and Juntunen, M. (2010). A comparison of WALS estimation with pretest and model selection alternatives with an application to costs of hip fracture treatments. In: Yamanishi, K. et al. (eds.), *Proceedings of the Third Workshop in Information Theoretic Methods in Science and Engineering WITMSE*, August 16-18, 2010, Tampere, Finland, TICSP Series.
4. Liski, A., Liski, E. P. and Häkkinen, U. (2013). Shrinkage estimation via penalized least squares in linear regression with an application to hip fracture treatment costs, *Proceedings of the 9th Tartu Conference on Multivariate Statistics*, World Scientific, Accepted for publication.
5. Liski, E. P. and Liski, A. (2008). MDL model averaging for linear regression. In: Grünwald, P. et al. (eds.) *Festschrift in honor of Jorma Rissanen on the occasion of his 75th birthday*, Tampere, Finland TICSP series.
6. Liski, A. and Liski, E. P. (2013). MDL model selection criterion for mixed models with an application to spline smoothing. In: SenGupta,

A., Samanta, T. & Basu, A. (eds.) *Statistical Paradigms - Recent Advances and Reconciliations*, Statistical Science and Interdisciplinary Research - Vol 14, World Scientific, Accepted for publication.

7. Liski, A., Tabus, I., Sund, R. and Häkkinen, U. (2012). Variable selection by sNML criterion in logistic regression with an application to a risk-adjustment model for hip fracture mortality, *Journal of Data Science*, 10(2), pp.321–343.

List of abbreviations

AIC	Akaike information criterion
BIC	Bayesian information criterion
BLUP	best linear unbiased predictor
LASSO	least absolute shrinkage and selection operator
LMM	linear mixed model
LS	least squares
MA	model averaging
MDL	minimum description length
ML	maximum likelihood
MS	model selection
MSE	mean square error
NML	normalized maximum likelihood
PenLS	penalized least squares
WALS	weighted average least squares
SC	stochastic complexity
SCAD	smoothly clipped absolute deviation
sNML	sequentially normalized maximum likelihood

1

Introduction

Much of modern scientific enterprise is concerned with the problem of model choice. A researcher collects data, usually in the form of measurements on many different characteristics of the observed units, and wants to explore the effects of these variables on some outcome of interest. This goal will be pursued by formulating a set of candidate models \mathcal{M} , say. Then we attempt to choose from \mathcal{M} a model that is a good explanation for the data. With a large number of models, it is clear that methods are needed to somehow summarize the qualities of the models under comparison. A major concern in model selection is overfitting: the selected model is overly complex. It fits well but predicts future data badly. Realizing that small changes in data may lead to a different model justifies the set-up of model averaging.

The thesis consists of an introductory part and of seven articles. The introductory part is divided into six chapters. The first chapter is an introduction to the concepts and ideas applied in the research articles, and the objective of the chapter is to give the big picture of the thesis. Chapter 2 gives an overview of the models and estimation methods covered in the seven research articles. Chapter 3 concerns model selection (MS) using the *minimum description length* principle (MDL), and finally Chapter 4 discusses certain applications of our proposed methodology on hip fracture mortality and hip fracture treatment cost data. Chapter 5 incorporates the summary of the attached research articles and highlights the author's contributions to them and the final chapter draws together the conclusions.

1.1 Statistical and information theoretic approaches

1.1.1 Information criteria

The information criteria have played a critical role in statistical modeling practice since Akaike (1973). In Akaike's approach the Kullback-Leibler

(1951) distance (see Subsection 3.1.2) is considered as the basic criterion for evaluating the goodness of a model as an approximation to the true distribution that generates the data. In practice, the Bayesian information criterion (BIC) proposed by Schwartz (1978) is also a widely used model selection criterion. The BIC is based on Bayesian probability and can be applied to models estimated by the ML method. The above methods are referred to as "information theoretic" because they utilize concepts from information theory. Mallows' (1973) criterion C_p and *generalized cross-validation* criterion GCV (Graven and Wahba 1979) are based on prediction error. We have applied the AIC, BIC and MDL criteria in the articles [5], [6] and [7]. In addition, we have used Mallows' C_p in [5], GCV in [6] and the c statistic (cf. Hosmer and Lemeshow 2000) as a diagnostic measure for logistic regression in [7]. The c statistic measures how well a model can discriminate between observations at different levels of the outcome.

Most MS methods are defined in terms of an appropriate information criterion (Claeskens and Hjort 2008). Let \mathcal{M} be the collection of models m which are considered as possible candidates for a final model. The general formula for Akaike's criterion, for example, is

$$AIC(m) = -2l(\hat{\beta}_m) + 2k_m,$$

where $\hat{\beta}_m$ is the ML estimate of a parameter β under a candidate model m , $l(\hat{\beta}_m)$ is the maximized log-likelihood and k_m is the length of the parameter vector β_m . Thus AIC is a penalized log-likelihood criterion which seeks for a balance between good fit and simplicity. The model \hat{m} with the lowest AIC score in \mathcal{M} is selected. The estimator of β may be represented as

$$\hat{\beta}_{AIC} = \sum_{m \in \mathcal{M}} I(m = \hat{m}) \hat{\beta}_m, \quad (1.1.1)$$

where \hat{m} is the model selected by AIC and $I(\cdot)$ is the indicator function with value 1 for the selected model and 0 for all other models. In general, the model selection probabilities $P(\hat{m} = m)$ depend on data and on the given MS procedure.

1.1.2 Minimum description length principle

The MDL principle has mainly been developed by Jorma Rissanen in a series of papers starting with the paper in 1978. It has its roots in the theory of Kolmogorov complexity (Li and Vitányi 1997). Kolmogorov's (1965) paper did serve as an inspiration for Rissanen's (1978) development of MDL. Another important inspiration for Rissanen was Akaike's (1973) AIC method for model selection, the first model selection method based on information theoretic ideas. This led to the development of the notion of stochastic complexity as the shortest codelength of the data given a model (Rissanen 1986 and 1987). However, the connection to Shtarkov's *normalized maximum likelihood* (NML) code was not made until 1996. An extensive introduction to the MDL history, philosophy and techniques can

be found in Grünwald (2007). MDL is also related to the Minimum Message Length Principle, developed by Wallace starting with the paper by Wallace and Boulton (1968).

Different authors may use 'MDL' in somewhat different meanings. It has sometimes been claimed that 'MDL = BIC'. Burnham and Anderson (2002, page 286) write "Rissanen's result is equivalent to BIC" and Hastie et al. (2001, p. 208) write 'MDL approach gives a selection criterion formally identical to BIC approach'. This is not quite true. However, under certain conditions Rissanen's 1978 criterion and BIC are asymptotically equivalent. Further, Hastie et al. (2001, p. 209) write 'the BIC criterion can also be viewed as a device for (approximate) model choice by minimum description length'. Miller (2002) refers to Rissanen's criteria presented in 1978 and 1987 papers. The latest formulation of the MDL principle can be found in Rissanen (2012, pp. 51–56). Our use of MDL is based on the ideas of NML and *stochastic complexity* (SC) introduced in Rissanen (1996).

The SC for data, relative to a suggested model, serves as a principal tool for model selection in this thesis. The computation of the SC can be considered as an implementation of the MDL principle. SC is the logarithm of the NML which contains two components: the maximized log likelihood and a component that may be interpreted as the parametric complexity of the model. If the parametric complexity of a model class is not bounded, there are alternative ways to deal with the problem (cf. Hansen and Yu 2001). We apply the MDL principle in linear Gaussian regression, in non-parametric regression by using *model averaging* (MA), in spline smoothing within the framework of linear mixed models (LMM) and in logistic regression. In Section 3.2 we introduce Rissanen's (2000) renormalization technique in Gaussian linear regression and illustrate its dependence on data constraints.

The introduction to MDL modeling in Section 3.1 reveals clearly the information theoretic roots of the MDL approach. The MDL implementation of Tabus and Rissanen (2006) for ordinary logistic regression is not computationally feasible in practice for such large data sets like our hip fracture data. Therefore, in [7] we implement sequential NML (sNML) for logistic regression which makes it possible to carry out MDL computations efficiently enough also for large data and models with large number of covariates.

1.1.3 Regression modeling

A successful application of statistical methods depends crucially on problem formulation where probability models have a central role. The choice of a family of possible models is thus a key step. Discussion on distinctions between different kinds of models is a substantial element of a research process when striving to clarify what to do in particular applications.

Likelihood-based regression models are important tools in data analysis. Typically a likelihood is assumed for a response variable y , and the mean or some other parameter is modeled as a linear function $\sum_{j=1}^p \beta_j x_j$ of a set of

explanatory variables x_1, \dots, x_p . The parameters of the linear function are then estimated by maximum likelihood. Examples of this are the normal linear regression model and the logistic regression model for binary data. Both of these models assume a linear form for the effects of explanatory variables. The Gaussian and logistic models are members of the class of generalized linear models (Nelder and Wedderburn 1972).

Linear regression

Regression analysis is one of the most often used tools in the statistician's toolbox. In articles [1] and [2] linear regression is the subject of research and in [3], [4] and [5] it is an important theoretical framework. The ordinary linear regression model (see e.g. Seber and Lee 2003) for response data y_i in relation to the values x_{i1}, \dots, x_{ip} of the p explanatory variables for individuals $i = 1, \dots, n$ is

$$y_i = x_{i1}\beta_1 + \dots + x_{ip}\beta_p + \varepsilon_i \quad \text{for } i = 1, \dots, n$$

with $\varepsilon_1, \dots, \varepsilon_n$ independently drawn from $N(0, \sigma^2)$ and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ a vector of regression coefficients. Typically $x_{i1} \equiv 1$, so that then β_1 is the intercept parameter. In matrix notation the model takes the form

$$\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n\}, \quad (1.1.2)$$

where \mathbf{X} is a $n \times p$ matrix of full rank with $n > p$.

Logistic regression

In article [7] we model hip fracture mortality with logistic regression. For each of the n patients we define $y_t = 1$ if the t th patient died within 90 days period after the hip fracture and $y_t = 0$ otherwise, and a corresponding model for the 365 days mortality is also considered. We treat the n binary outcome variables y_1, \dots, y_n as independent. Let

$$\pi(\mathbf{x}_t; \boldsymbol{\beta}) = P(y_t = 1), \quad t = 1, \dots, n,$$

and assume that

$$\log \frac{\pi(\mathbf{x}_t; \boldsymbol{\beta})}{1 - \pi(\mathbf{x}_t; \boldsymbol{\beta})} = \boldsymbol{\beta}' \mathbf{x}_t,$$

where $\mathbf{x}_t = (x_{t1}, \dots, x_{tk})'$ is the vector of k covariate values of the t th patient. Since the y_1, \dots, y_n are independent and Bernoulli distributed, the likelihood function of $\boldsymbol{\beta}$ is

$$L(\boldsymbol{\beta}; \mathbf{x}_1, \dots, \mathbf{x}_n) = \prod_{t=1}^n \pi(\mathbf{x}_t; \boldsymbol{\beta})^{y_t} [1 - \pi(\mathbf{x}_t; \boldsymbol{\beta})]^{1-y_t}. \quad (1.1.3)$$

In this context we introduce a sequential NML (sNML) approach to selection of covariates.

Nonparametric regression

A tendency to move away from linear functions and to model the dependence of y on x_1, \dots, x_p in a more nonparametric fashion has gained strength in the past few years. For a single explanatory variable, such a model would be

$$y = \mu(x) + \varepsilon,$$

where ε is the error term and $\mu(x)$ an unspecified smooth function. This function can be estimated by any so-called scatterplot smoother, for example a running mean, running median, running least squares line, kernel estimate or a spline (see Ruppert *et al.* 2003 for discussions of smoothing techniques). For the p covariates x_1, \dots, x_p , one can use a p -dimensional scatterplot smoother to estimate $\mu(x)$, or a combination of a parametric and nonparametric model.

In [5] we consider modeling in nonparametric regression by utilizing linear models and MA. We assume that the mean function belongs to a function class (infinite dimensional) whose elements admit representations as infinite dimensional linear models. The practical significance of such models is that they may be well approximated by a finite number of leading terms. To obtain an estimate of the mean function we employ a set of nested approximating linear models. We construct smooth estimators of the mean function across approximating linear models using MA.

Smoothing spline models are known for their flexibility in fitting a mean function on a given index set. We propose in [6] a spline smoothing technique that combines the power of a smoothing spline model and a linear mixed model (LMM). For spline smoothing we rewrite the smooth estimation as a LMM. Smoothing methods that use basis functions with penalization can utilize the likelihood theory in the LMM framework. We introduce the NML model selection criterion for the LMM and propose an automatic data-based spline smoothing method that utilizes MDL model selection. We compare the performance of the MDL method in simulation experiments with three alternatives which use AIC, BIC and GCV model selection methods.

1.2 Selection or averaging?

1.2.1 The traditional t-test

Traditionally the most widely used MS method in multiple regression is to carry out a sequence of tests in order to find out the nonzero regression coefficients and to select the corresponding regressors. We consider the following enlarged model

$$\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta} + z\gamma, \sigma^2\mathbf{I}\}, \quad (1.2.4)$$

where the regressor z is added to the model (1.1.2). Here σ^2 is assumed known. It is unrealistic, but simplifies the presentation of main ideas. In

practical applications, this restriction is relaxed, of course. Following Magnus et al. (2010) the x-variables are called "focus" regressors and the z-variable "auxiliary" regressor. We distinguish between variables because the focus regressors we want to keep in the model on theoretical or other grounds whereas the auxiliary regressor is added to the model only if it is supposed to improve estimation of the coefficients of the focus regressors.

Let \mathcal{M}_0 denote the model (1.1.2) and \mathcal{M}_1 model (1.2.4). The least squares (LS) estimator of β in \mathcal{M}_0 is

$$\hat{\beta}_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Using the notation $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$,

$$\mathbf{q} = \frac{\sigma}{\sqrt{\mathbf{z}'\mathbf{M}\mathbf{z}}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z} \quad \text{and} \quad \theta = \frac{\gamma}{\sigma/\sqrt{\mathbf{z}'\mathbf{M}\mathbf{z}}}, \quad (1.2.5)$$

we can write the LS estimates of β and γ in \mathcal{M}_1 as

$$\hat{\beta}_1 = \hat{\beta}_0 - \hat{\theta}\mathbf{q}, \quad \hat{\gamma} = \frac{\mathbf{z}'\mathbf{M}\mathbf{y}}{\mathbf{z}'\mathbf{M}\mathbf{z}},$$

where $\hat{\theta} = \frac{\hat{\gamma}}{\sigma} \sqrt{\mathbf{z}'\mathbf{M}\mathbf{z}}$ is called the t -ratio, which follows the normal distribution $\mathbf{N}(\theta, 1)$. Note that $\hat{\theta}$ and $\hat{\beta}_0$ are independent.

Which of the two models \mathcal{M}_0 and \mathcal{M}_1 should we use to estimate β ? The traditional statistical practice is to decide between the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ by testing the hypothesis $\theta = 0$ against $\theta \neq 0$ (equivalently $\gamma = 0$ against $\gamma \neq 0$). If the t -ratio is large, the hypothesis is rejected and the model \mathcal{M}_1 is selected. This implies that we use the estimate $\hat{\beta}_1$. Otherwise we select the model \mathcal{M}_0 and use the estimate $\hat{\beta}_0$. This is model selection between \mathcal{M}_0 and \mathcal{M}_1 . In fact, the choice between $\hat{\beta}_0$ and $\hat{\beta}_1$ is the estimator

$$\tilde{\beta} = \begin{cases} \hat{\beta}_0 & \text{if } |\hat{\theta}| \leq c; \\ \hat{\beta}_1 & \text{if } |\hat{\theta}| > c, \end{cases} \quad (1.2.6)$$

for some nonnegative constant c . The value of c corresponds to the significance level of the test, e.g. $c = 1.96$ corresponds to the 5% level.

If we are interested in best possible estimation of β , not γ , testing the hypothesis $\gamma = 0$ may not be the correct procedure. Toro-Vizcarrondo and Wallace (1968) developed a test which makes it possible to compare estimators with respect to their mean square error (MSE). Then the question is: "Is $\hat{\beta}_0$ a better estimator of β than $\hat{\beta}_1$ with respect to their MSE criterion?" But it turns out that also the Toro-Vizcarrondo and Wallace test leads to an estimator of the form (1.2.6). A general theory of comparing estimators with respect to the MSE criterion can be found e.g. in Rao et al. (2008) or in Judge and Bock (1978).

1.2.2 Weighted average

The estimator (1.2.6) is actually of the form

$$\tilde{\beta} = \lambda(\hat{\theta})\hat{\beta}_1 + [1 - \lambda(\hat{\theta})]\hat{\beta}_0, \quad 0 \leq \lambda(\hat{\theta}) \leq 1, \quad (1.2.7)$$

where $\lambda(\hat{\theta})$ is the indicator function $\lambda(|\hat{\theta}| > c)$. It takes the value 1 if $|\hat{\theta}| > c$, and otherwise 0. Following Magnus (1999) we consider a more general class of weighting functions than just indicators. The estimator (1.2.7) is called the weighted average least squares estimator (WALS) of β if the real-valued function λ of $\hat{\theta}$ satisfies certain regularity conditions. Usually λ is a nondecreasing function of $|\hat{\theta}|$, so that the larger $|\hat{\theta}|$, the larger λ will be and hence more weight will be put on $\hat{\beta}_1$ relative to $\hat{\beta}_0$. This is *model averaging*. MA can be viewed as a two-step procedure.

1. Estimate β conditional upon the selected models \mathcal{M}_0 and \mathcal{M}_1 .
2. Compute the estimate of β as a weighted average of these conditional estimates.

The following equivalence theorem, originally proved by Magnus and Durbin (1999), and later extended by Danilov and Magnus (2004) turns out useful in the study of WALS estimators.

Equivalence theorem. *Let $\tilde{\beta} = \lambda(\hat{\theta})\hat{\beta}_1 + [1 - \lambda(\hat{\theta})]\hat{\beta}_0$ be a WALS estimator of β and $\tilde{\theta} = \lambda(\hat{\theta})\hat{\theta}$, where $\lambda(\hat{\theta})$ is as in (1.2.7) and $\hat{\theta} \sim \mathbf{N}(\theta, 1)$. Then*

$$\text{MSE}(\tilde{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \text{MSE}(\tilde{\theta})\mathbf{q}\mathbf{q}',$$

where \mathbf{q} is defined in (1.2.5) and $\text{MSE}(\cdot)$ is the mean square error of an estimator.

The equivalence theorem expresses the MSE of a WALS estimator $\tilde{\beta}$ of β as a function of the MSE of the estimator $\tilde{\theta}$ of θ . Thus $\text{MSE}(\tilde{\beta})$ is minimized if and only if $\text{MSE}(\tilde{\theta})$ is minimized. A more general version of the equivalence theorem is considered in articles [3] and [4]. The equivalence theorem is important because it shows that finding the best WALS estimator of β is equivalent to finding the best estimator

$$\tilde{\theta} = \lambda(\hat{\theta})\hat{\theta} \quad (1.2.8)$$

of θ for the simple normal distribution $N(\theta, 1)$. Thus the problem is to find a λ -function which would yield a good estimator of θ with respect to its mean square error

$$\begin{aligned} \text{MSE}(\tilde{\theta}; \theta) &= \mathbf{E}(\tilde{\theta} - \theta)^2 \\ &= \text{Var}(\tilde{\theta}) + [\mathbf{E}(\tilde{\theta}) - \theta]^2. \end{aligned}$$

The traditional t-test for testing $\theta = 0$ against $\theta \neq 0$ yields an estimator which is obtained from (1.2.8) by choosing $\lambda(\hat{\theta}) = \lambda(|\hat{\theta}| > c)$:

$$\tilde{\theta}_H = \lambda(|\hat{\theta}| > c)\hat{\theta} \quad (1.2.9)$$

Donoho and Johnstone (1994) called the estimator (1.2.9) the hard shrinkage function in the context of wavelet shrinkage denoising. Its MSE is

$$\text{MSE}(\tilde{\theta}_H; \theta) = \theta + \theta[1 - \Phi(c - \theta) - \Phi(c + \theta)] + \varphi(c - \theta) - \varphi(c + \theta),$$

where $\Phi(\cdot)$ is the standard Gaussian distribution function and $\varphi(\cdot)$ is the standard Gaussian density function. Magnus (1999) showed that (1.2.9) has many undesirable properties with regard to its MSE and the WALS estimators have advantages (cf. Magnus 2002) over the pretest estimators (1.2.7). In article [4] we have defined a class of shrinkage estimators in view of finding estimators which have uniformly low MSE with respect to the theoretical minimum which we have derived.

1.2.3 Least squares model averaging

In the model (1.2.4) there is only one auxiliary regressor. This yields two candidate models: \mathcal{M}_0 and \mathcal{M}_1 . In articles [3] and [4] we consider a more general enlarged model which contains m auxiliary regressor. A candidate model is constructed by selecting a subset of z -variables to the model where x -variables are kept fixed. Thus the number of candidate models is 2^m . Let $\hat{\beta}_i$ denote the LS estimator of β from the candidate model \mathcal{M}_i when the models are properly indexed, and let $\lambda = (\lambda_0, \lambda_1, \dots, \lambda_M)'$ denote a vector of nonnegative weights which sum to one. Then the weighted average of the LS estimators takes the form

$$\tilde{\beta} = \sum_{i=0}^M \lambda_i \hat{\beta}_i, \quad (1.2.10)$$

where $M = 2^m$. Magnus et al. (2010) called the estimator (1.2.10) weighted average LS (WALS) estimator, if it satisfies the following minimal regularity conditions:

$$\lambda_i \geq 0, \quad \sum_i \lambda_i = 1 \quad \text{and} \quad \lambda_i = \lambda_i(\mathbf{M}\mathbf{y}). \quad (1.2.11)$$

So, the last condition in (1.2.11) means that the weights may depend on the least squares residuals. They also proposed an estimation technique which avoids estimation of the single model weights. Hansen (2007) showed that a LS model averaging estimator like (1.2.10) can achieve lower MSE than any individual LS estimator $\hat{\beta}_i$.

According to Magnus et al. (2010) maybe Leamer (1978) was the first to categorize variables into two classes, which he called 'focus' and 'doubtful' variables. The focus variables are always in the model, while the doubtful variables can be combined in an arbitrary linear manner. Consequently,

exclusion of any subset of doubtful variables is a special case of this linear combination. Note that the focus variables are not always the focus of the study although they are always in the model. In the analysis of covariance model, variables are also categorized into two classes: x -variables and z -variables. Cox and McCullagh (1982) outline six different aspects of analysis of covariance. Our application in [3] comes close to the aspect which they call 'Adjustment for bias in observational studies'.

In [3] we apply the WALS technique to compare the hip fracture treatment costs of the seven largest hospital districts in Finland. We are interested in differences of treatment costs, therefore hospital districts are the focus variables. The set of 38 auxiliary variables contains important comorbidities like congestive heart failure, diabetes and cancer, for example. The patients are not randomly allocated to hospitals, and therefore the patient case mix may vary considerably between hospital districts. The auxiliary regressors are intended to adjust for bias due to possible differences in patient case mix and to improve the precision of comparisons. In article [4] we introduce for WALS a penalized LS estimation technique which avoids estimation of single model weights. Consequently, the technique is computationally very efficient. Estimators are evaluated with respect to their MSE. So, our approach to WALS is based on traditional statistical methods.

Magnus et al. (2010) applied WALS to growth empirics, they utilized the research of growth models in economics to select the focus regressors and auxiliary regressors to their growth model. Further, they analysed the same growth data using also Bayesian model averaging (BMA) and compared the two averaging methods: BMA and WALS. WALS had two major advantages over BMA: its computational burden is trivial and it is based on transparent definition of prior ignorance. Magnus et al. (2010) derive the advocated shrinkage function using the Laplace prior. Thus their WALS approach is a Bayesian method although not standard BMA. Einmahl et al. (2011) introduced the Subbotin prior which they claimed to be 'suitable'.

The main theme of article [5] is the problem of selecting the weights for averaging across a set of approximating linear models. Buckland et al. (1997) proposed weights proportional to $\exp(-AIC_m/2)$, where AIC_m is the AIC score for a model m . Similar weighting can be derived from other model selection criteria as well. In distinction, Hansen (2007) proposed selecting the weights by minimizing the Mallows' criterion. We present an MDL based solution to the weight selection problem. This approach is suitable in applications where the number of candidate models is not very large as in [5]. We compare the performance of the MDL method with the performance of the above mentioned alternative weight selection methods in simulation experiments.

2

Estimation and averaging

Our framework in articles [3] and [4] is an enlarged linear regression model

$$\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}, \sigma^2 \mathbf{I}_n\}, \quad (2.0.1)$$

where \mathbf{X} is an $n \times p$ matrix of focus variables as in the model (1.1.2) and \mathbf{Z} is an $n \times m$ matrix containing m additional explanatory variables which are called the auxiliary variables. The model (2.0.1) which includes all m auxiliary variables is called the unrestricted model. The model (1.1.2) is the fully restricted case of (2.0.1), obtained by putting the restriction $\boldsymbol{\gamma} = \mathbf{0}$. In Subsection 1.2.1 we have considered the case $m = 1$, and in Subsection 1.2.3 we have also discussed, in view of applications, the interpretation of categorizing variables into two classes. The whole family of restricted models and their indexing is introduced in Subsection 2.1.1. The matrix (\mathbf{X}, \mathbf{Z}) is assumed to be of full column rank.

A common approach is to test the hypothesis $\boldsymbol{\gamma} = \mathbf{0}$ against $\boldsymbol{\gamma} \neq \mathbf{0}$ and to include \mathbf{Z} into the model if the hypothesis $\boldsymbol{\gamma} = \mathbf{0}$ is rejected and exclude the z -variables otherwise. Then the alternative estimators of $\boldsymbol{\beta}$ are the restricted LS estimator under the restriction $\boldsymbol{\gamma} = \mathbf{0}$, i.e. estimation in the model (1.1.2), say \mathcal{M}_0 , and the unrestricted LS estimator in the model (2.0.1), say \mathcal{M}_M . The relative performance of the estimators may be assessed by the mean squared error criterion (MSE).

2.1 Least squares estimation

The LS estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ under the model \mathcal{M}_M are (Seber 1977, p. 66 and Seber and Lee 2003, p. 54)

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \hat{\boldsymbol{\beta}}_0 - \dot{\mathbf{Q}}\hat{\boldsymbol{\gamma}} \\ \hat{\boldsymbol{\gamma}} &= (\mathbf{Z}'\mathbf{M}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{M}\mathbf{y}, \end{aligned}$$

respectively, where $\hat{\beta}_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is the LS estimator of β from the fully restricted model \mathcal{M}_0 , $\hat{\mathbf{Q}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}$ and

$$\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}. \quad (2.1.2)$$

The matrix \mathbf{X}' denotes the transpose of \mathbf{X} and $(\mathbf{X}'\mathbf{X})^{-1}$ is the inverse of $\mathbf{X}'\mathbf{X}$.

It will be convenient to work with the canonical form of the model (2.0.1) where z -variables are orthogonalized. In article [4] we derive a new (α, θ) -parametrization, where parameter $\alpha = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\gamma$ and $\theta = \mathbf{C}^{-1}\gamma$. A nonsingular matrix \mathbf{C} can be chosen such that $\mathbf{C}'\mathbf{Z}'\mathbf{M}\mathbf{Z}\mathbf{C} = \mathbf{I}_m$ (Horn and Johnson 1985, p. 466), since $(\mathbf{M}\mathbf{Z})'\mathbf{M}\mathbf{Z} = \mathbf{Z}'\mathbf{M}\mathbf{Z}$ is positive definite (Seber 1977, p. 66). There exists one-to-one correspondence between (β, γ) - and (α, θ) -parametrizations, and consequently the same correspondence holds also between their LS estimates.

We may now write the unrestricted model \mathcal{M}_M in (2.0.1) using (α, θ) -parametrization as follows

$$\mathcal{M}_M : \{\mathbf{y}, \mathbf{X}\alpha + \mathbf{U}\theta, \sigma^2\mathbf{I}_n\}, \quad (2.1.3)$$

where $\mathbf{U} = \mathbf{M}\mathbf{Z}\mathbf{C}$ denotes the matrix of orthogonal canonical auxiliary regressors. In (α, θ) -parametrization we have

$$\begin{aligned} \alpha &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{C}\mathbf{C}^{-1}\gamma \\ &= \beta + \mathbf{Q}\theta, \end{aligned}$$

where $\mathbf{Q} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{C}$. The model \mathcal{M}_M is orthogonal such that $\mathbf{X}'\mathbf{U} = \mathbf{0}$ and (\mathbf{X}, \mathbf{U}) is of full column rank. Then the LS estimators of α and θ from the model \mathcal{M}_M are

$$\begin{aligned} \hat{\alpha} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \equiv \hat{\beta}_0, \\ \hat{\theta} &= \mathbf{U}'\mathbf{y}. \end{aligned}$$

The correspondence between the vectors (α', θ') and (β', θ') is one-to-one, and consequently the same correspondence holds between their LS estimates. Because of the one-to-one correspondence between the two parametrizations the LS estimator of β under the unrestricted model \mathcal{M}_M is (cf. Seber 1977, p. 66 and Seber and Lee 2003, p. 54)

$$\begin{aligned} \hat{\beta}_M &= \hat{\alpha} - \mathbf{Q}\hat{\theta} \\ &= \hat{\beta}_0 - \hat{\mathbf{Q}}\hat{\gamma}. \end{aligned}$$

2.1.1 Restricted least squares

In the unrestricted model \mathcal{M}_M in (2.1.3) there are m components of θ , and 2^m submodels are obtained by setting various subsets of the elements $\theta_1, \dots, \theta_m$ of θ equal to zero. These 2^m models $\mathcal{M}_0, \dots, \mathcal{M}_M$ can be written as

$$\mathcal{M}_i : \{\mathbf{y}, \mathbf{X}\alpha + \mathbf{U}_i\theta, \sigma^2\mathbf{I}_n\}, \quad (2.1.4)$$

where $\mathbf{U}_i = \mathbf{U}\mathbf{W}_i$ and $\mathbf{W}_i = \text{diag}(w_{i1}, \dots, w_{im})$, $i = 0, 1, \dots, M$ are $m \times m$ diagonal matrices with diagonal elements $w_{ij} \in \{0, 1\}$, $j = 1, \dots, m$, and $M = 2^m - 1$. In articles [3] and [4] we seek a smooth compromise across the set of competing models $\mathcal{M}_0, \dots, \mathcal{M}_M$ with model averaging (MA). The weighted average LS (WALS) estimation is the topic of [3] and WALS was a motivating impetus also in [4].

We may suppose that the models are in increasing order with respect to diagonal elements of \mathbf{W}_i when the diagonals are interpreted as m -digit binary numbers $w_{i1} \dots w_{im}$, $i = 0, 1, \dots, M$. Then the indices $1, \dots, M$ are associated with the diagonals as follows

$$\begin{aligned} 0 &\rightarrow 00 \dots 0, & 1 &\rightarrow 0 \dots 01, & 2 &\rightarrow 0 \dots 010, & 3 &\rightarrow 0 \dots 011, \dots, \\ M-1 &\rightarrow 11 \dots 10, & M &\rightarrow 11 \dots 11, \end{aligned}$$

where the number of models is $M + 1 = 2^m$. Standard theory of LS estimation with linear restrictions (Seber 1977 and Seber and Lee 2003) yields the restricted LS estimators

$$\hat{\beta}_i = \hat{\beta}_0 - \mathbf{Q}\mathbf{W}_i\hat{\theta} \quad (2.1.5)$$

for β under the models \mathcal{M}_i , $0 \leq i \leq M$.

2.1.2 Averaging across restricted LS estimators

A LS model averaging estimator of β (cf. (1.2.10)) is obtained by taking a weighted average of the LS estimators (2.1.5) which takes the form

$$\begin{aligned} \tilde{\beta} &= \sum_{i=0}^M \lambda_i \hat{\beta}_i = \sum_{i=0}^M \lambda_i (\hat{\beta}_0 - \mathbf{Q}\mathbf{W}_i\hat{\theta}) \\ &= \hat{\beta}_0 - \mathbf{Q}\mathbf{W}\hat{\theta}, \end{aligned} \quad (2.1.6)$$

where $\mathbf{W} = \sum_{i=0}^M \lambda_i \mathbf{W}_i$. The estimator (2.1.6) is a WALS estimator if it satisfies the regularity conditions (1.2.11). In practice the number of weights to be estimated may be huge. For example in article [3] the number of z -variables is 38, and consequently the number of candidate models is 2^{38} . Therefore the set of candidate models is usually restricted to a small fraction of all possible models.

However, the effect of this 'preselection' may be difficult to assess. Therefore, we have studied WALS which avoids estimation of the single model weights. This feature makes it computationally attractive. In article [3] our goal is to accomplish the positive features of WALS by using shrinkage estimation. In article [4] we approach WALS and shrinkage estimation by using a two-step LS. Then weight estimation can be carried out by using penalized LS (PenLS) estimation without heavy computational burden (see Subsection 2.1). This approach provides new insight into weight estimation providing a variety of alternative estimators with good risk properties.

2.1.3 Shrinking LS estimates

By the equivalence theorem of Danilov and Magnus (2004, Theorem 1) the important statistical properties of the WALS estimator (2.1.6) depend only on the estimator

$$\tilde{\theta} = \mathbf{W}\hat{\theta}, \quad 0 \leq |\tilde{\theta}_i| \leq |\hat{\theta}_i|, \quad i = 1, \dots, m, \quad (2.1.7)$$

of θ , where $\hat{\theta}$ is the LS estimator of θ and \mathbf{W} is an $m \times m$ diagonal matrix with diagonal elements w_i , $0 \leq w_i \leq 1$, $i = 1, \dots, m$. Thus w_i 's shrink the LS estimates $\hat{\theta}_i$ towards zero, and consequently $0 \leq |\tilde{\theta}_i| \leq |\hat{\theta}_i|$, $i = 1, \dots, m$. We posit that each diagonal element $w_i = w_i(\hat{\theta}_i)$ depends on data such that w_i is a function of only the i th element $\hat{\theta}_i$ of $\hat{\theta}$, $i = 1, \dots, m$. Further, we assume that the shrinkage functions are even: $w_i(-\hat{\theta}_i) = w_i(\hat{\theta}_i)$, $i = 1, \dots, m$. Thus the functions $\tilde{\theta}_i$ are odd: $\tilde{\theta}_i(-\hat{\theta}_i) = -\tilde{\theta}_i(\hat{\theta}_i)$. Magnus *et al.* (2010) and Einmahl *et al.* (2011) adopted a Bayesian view on estimation deciding on to advocate the Laplace and Subbotin estimators which are of shrinkage type.

The trick in our approach to MA is to convert estimation of the model weights into estimation of the shrinkage factors w_i , $i = 1, \dots, m$. The number of shrinkage factors increase linearly with the number of explanatory variables in regression whereas the number of candidate models, and the number of model weights respectively, increase exponentially. When estimation of a weight for every single model $\mathcal{M}_0, \dots, \mathcal{M}_M$ in (2.1.4) is required, the computing time will be of order 2^m . If the number of auxiliary regressors is large in (2.0.1), say $m = 50$, then computing time will be of order 2^{50} which is infeasible. Thus the proposed MA technique is computationally superior to techniques that require estimation of every single weight.

This motivates us to study shrinkage estimators. We will now define an important class of estimators for θ which we call *shrinkage estimators*, and in the sequel we denote it by \mathcal{S} .

Definition 1. A real valued estimator δ of θ defined on \mathbb{R} is a shrinkage estimator if the following four conditions hold:

- (a) $0 \leq \delta(\hat{\theta}) \leq \hat{\theta}$ for $\hat{\theta} \geq 0$,
- (b) $\delta(-\hat{\theta}) = -\delta(\hat{\theta})$,
- (c) $\delta(\hat{\theta})/\hat{\theta}$ is nondecreasing on $[0, \infty)$ and
- (d) $\delta(\hat{\theta})$ is continuous,

where $\hat{\theta}$ is the LS estimator of θ .

In addition to shrinkage (a) and antisymmetry (b) properties, the definition puts two further requirements for shrinkage estimators. Consider now the condition (c). Denote $w(\hat{\theta}) = \delta(\hat{\theta})/\hat{\theta}$ for $\hat{\theta} > 0$ and think $\delta(\hat{\theta})$ as a weighted average of $\hat{\theta}$ and 0: $\delta(\hat{\theta}) = w(\hat{\theta})\hat{\theta} + (1 - w(\hat{\theta}))0$. The larger is $|\hat{\theta}|$,

the better $\hat{\theta}$ is as an estimator of θ . Hence, when $\hat{\theta}$ increases we wish to put more weight on $\hat{\theta}$ than on 0, i.e., we wish to make $w(\hat{\theta})$ larger. Thus we see that the condition (c) makes sense. Condition (d) is a minimal smoothness condition which guarantees certain stability of estimation in the sense that small changes of data cannot create excessive variation of estimates.

There exists an extensive statistical literature on shrinkage estimation. Perhaps the most famous shrinkage estimator is the James-Stein estimator (1961). Another long-time favourite is the ridge estimator of Hoerl and Kennard (1970). Shrinkage estimators are continuously a topic of active research, among them are for example LASSO (Tibshirani 1996) and non-negative garrote (Breiman 1995). Usually an estimator is called shrinkage estimator if it has the shrinkage property (a). However, we call the estimators in \mathcal{S} shrinkage estimators, although \mathcal{S} is only a subclass of estimators which have the shrinkage property (a). So, in general the elements $\hat{\theta}_i$, $i = 1, \dots, m$ of $\hat{\theta}$ in (2.1.7) are not in \mathcal{S} . Even the ridge estimator is not a shrinkage estimator in the sense of Definition 1, as is shown in Example 2.1.2 and Example 2.1.1.

We are also restricted to the real valued estimators, since in estimation of θ we finally need to solve m one-dimensional estimation problems. One motivation of Definition 1 is that the estimators \mathcal{S} have the efficiency bound (2.1.9) (cf. Theorem 4.1, article [4]). Further advantage of \mathcal{S} is that it contains many well known estimators with a desirable risk profile, see the definition below in connection of the regret (2.1.10). The idea of shrinkage estimation is quite general and it has been applied widely, also outside the world of linear models. Gruber (1998) provides the basic theory and surveys the extensive literature so far.

In estimation of θ in (2.1.11) we will use the penalized LS technique. If the penalty function satisfies proper regularity conditions, then the penalized LS yields a solution which is a shrinkage estimator of θ . In this approach we choose a suitable penalty function in order to get a shrinkage estimator with good risk properties. So, we are able to characterize a variety of interesting estimators from which many have already shown their potential in applications. This technique is also computationally efficient. The related Bayesian technique is to impose certain restrictions on the prior density, see e.g. Einmahl et al. (2011).

We prove in [4] Theorem 4.1 that gives sufficient conditions for the PenLS estimate of θ to be a shrinkage estimator. Further, the theorem provides the lower bound of the mean squared error

$$\text{MSE}[\delta(\hat{\theta}); \theta] = \mathbf{E}[\delta(\hat{\theta}) - \theta]^2 = \text{Var}[\delta(\hat{\theta})] + \text{Bias}[\delta(\hat{\theta}); \theta] \quad (2.1.8)$$

of $\delta(\hat{\theta})$, where $\text{Bias}[\theta, \delta(\hat{\theta})] = \{\mathbf{E}[\delta(\hat{\theta})] - \theta\}^2$. This lower bound

$$\inf_{\delta(\hat{\theta}) \in \mathcal{S}} \text{MSE}(\delta(\hat{\theta}); \theta) = \frac{\theta^2}{1 + \theta^2}. \quad (2.1.9)$$

is called the *efficiency bound* of $\delta(\hat{\theta})$.

Note that the hard thresholding (pretest) estimator $\tilde{\theta}_H$ given in (1.2.9) is not continuous, and hence it does not belong to the class of shrinkage estimators \mathcal{S} . Magnus (1999) demonstrates a number of undesirable properties of the pretest estimator. It is inadmissible and there is a range of values for which the MSE of $\tilde{\theta}_H$ is greater than the MSE of both the least squares estimator $\hat{\theta}(z) = z$ and the null estimator $\hat{\theta}(z) \equiv 0$. The traditional pretest at the usual 5% level of significance results in an estimator that is close to having worst possible performance with respect to the MSE criterion in the neighborhood of the value $|\theta/\sigma| = 1$ which was shown to be of crucial importance.

Example 2.1.1. The L_q penalty $p_\lambda(|\theta|) = \lambda |\theta|^q$, $q \geq 0$ results in a bridge regression (Frank and Friedman, 1993). The derivative $p'_\lambda(\cdot)$ of the L_q penalty is nonincreasing on $[0, \infty)$ only when $q \leq 1$ and the solution is continuous only when $q \geq 1$. Therefore, only L_1 penalty in this family yields a shrinkage estimator. This estimator is a soft thresholding rule, proposed by Donoho and Johnstone (1994),

$$\tilde{\theta}_S = \text{sgn}(z)(|z| - \lambda)_+,$$

where z_+ is shorthand for $\max\{z, 0\}$. LASSO (Tibshirani, 1996) is the PenLS estimate with the L_1 penalty in the general least squares and likelihood settings.

Since we have the efficiency bound of the shrinkage estimators $\delta(\hat{\theta})$, the *regret* of $\delta(\hat{\theta})$ can be defined as

$$r[\delta(\hat{\theta}); \theta] = \text{MSE}[\delta(\hat{\theta}); \theta] - \frac{\theta^2}{1 + \theta^2}. \quad (2.1.10)$$

We wish to find an estimator with the desirable property that its risk (MSE) is uniformly close to the infeasible efficiency bound. If we know the distribution of $\delta(\hat{\theta})$, we can determine the risk $\text{MSE}[\delta(\hat{\theta}); \theta]$ and regret $r[\delta(\hat{\theta}); \theta]$ curves of an estimator $\delta(\hat{\theta})$ as a function of θ . These functions define *the risk profile* of $\delta(\hat{\theta})$.

In theoretical considerations σ^2 is assumed to be known, and hence we can always consider the variable z/σ when $\text{Var}(z) = \sigma^2$. Then, under the normality assumption, the expectation is simply taken with respect to the normal distribution $N(\theta, 1)$, and comparison of estimators risk performance is done under this assumption. A typical technique is to consider the risk (2.1.8) or the regret (2.1.10) as a function of θ .

Although we have not explicitly displayed results of risk comparisons between estimators, such procedure is carried out when implementing the various estimators into practice. Such comparisons are also available in literature. Let us consider, for example, the SCAD estimator that was applied in [4]. It includes two tuning parameters whose values must be fixed to make the estimator usable. The aim is to fix the values of the tuning parameters so that a favourable risk profile is obtained. In practical

applications we replace the unknown σ^2 with s^2 , the estimate of σ^2 in the unrestricted model. Danilov and Magnus (2004) demonstrated that effects of estimating σ^2 are small. They used the Laplace estimator as a shrinkage function. We expect the approximation to be accurate for other shrinkage estimators too, although more work is needed to clarify this issue.

2.1.4 Penalized least squares

Fitting the orthogonalized model (2.1.3) can be considered as a two-step least squares procedure (Seber 1977). The first step is to calculate $\hat{\beta}_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and replace \mathbf{y} by $\mathbf{y} - \mathbf{X}\hat{\beta}_0 = \mathbf{M}\mathbf{y}$, where \mathbf{M} is defined in (2.1.2). Then denote $\mathbf{z} = \mathbf{U}'\mathbf{y}$, and note that from the definition of \mathbf{U} below (2.1.3) follows the equality $\mathbf{U}'\mathbf{M} = \mathbf{U}'$. Then the model \mathcal{M}_M in (2.1.3) takes the form

$$\mathbf{z} = \boldsymbol{\theta} + \mathbf{U}'\boldsymbol{\varepsilon}, \quad \mathbf{U}'\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\mathbf{I}_m). \quad (2.1.11)$$

The second step is to estimate $\boldsymbol{\theta}$ from the model (2.1.11).

In [4] we have carried out estimation of $\boldsymbol{\theta}$ by applying the penalized LS technique. If the penalty function satisfies proper regularity conditions, then the penalized LS yields a solution which is a shrinkage estimator of $\boldsymbol{\theta}$. In this approach we choose a suitable penalty function in order to get a shrinkage estimator with good risk properties. The related Bayesian technique is to impose certain restrictions on the prior density, see e.g. Einmahl *et al.* (2011). So, we are able to characterize a variety of interesting estimators from which many have already shown their potential in applications. This technique is also computationally efficient.

The penalized least squares estimate (PenLS) of $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)'$ is the minimizer of

$$\frac{1}{2} \sum_{i=1}^m (z_i - \theta_i)^2 + \sum_{i=1}^m p_\lambda(|\theta_i|), \quad (2.1.12)$$

where $\lambda > 0$. It is assumed that the penalty function $p_\lambda(\cdot)$ is

- (i) nonnegative,
- (ii) nondecreasing and
- (iii) differentiable on $[0, \infty)$.

Minimization of (2.1.12) is equivalent to minimization componentwise. Thus we may simply minimize

$$l(\theta) = \frac{1}{2}(z - \theta)^2 + p_\lambda(|\theta|)$$

with respect to θ .

Example 2.1.2. There are close connections between the PenLS and variable selection or the PenLS and ridge regression, for example. Taking the

L_2 penalty $p_\lambda(|\theta|) = \frac{\lambda}{2}|\theta|^2$ yields the ridge estimator

$$\check{\theta}_R = \frac{1}{1 + \lambda} z,$$

where $\lambda > 0$. The hard thresholding penalty function

$$p_\lambda(|\theta|) = \lambda^2 - \frac{1}{2}(|\theta| - \lambda)^2 \mathbb{I}(|\theta| < \lambda)$$

yields the hard thresholding rule

$$\check{\theta}_H = z \{ \mathbb{I}(|z| > \lambda) \},$$

where $\mathbb{I}(\cdot)$ is the indicator function. Then the minimizer of the expression (2.1.12) is $z_j \{ \mathbb{I}(|\theta_j| > \lambda) \}$, $j = 1, \dots, m$, and it coincides with the best subset selection for orthonormal designs. In statistics (see e.g. Morris *et al.* 1972) and in econometrics (see, e.g. Judge *et al.* 1985), the hard thresholding rule is traditionally called the pretest estimator.

2.2 Maximum likelihood estimation

2.2.1 MLE in linear mixed model

In article [6] we consider model selection for linear mixed models (LMM) using the NML criterion (Rissanen 1996). Regression splines that use basis functions with penalization can be fitted conveniently using the machinery of LMM's, and thereby borrow from a rich source of existing methodology (cf. Brumback *et al.* 1999 and Ruppert *et al.* 2003). The basis coefficients can be considered as random coefficients and the smoothing parameter as the ratio between variances of the error variables and random effects, respectively.

We posit the smoothing model

$$y_i = r(x_i) + \sigma \varepsilon_i, \quad i = 1, \dots, n,$$

where $r(\cdot)$ is a smooth function giving the conditional mean of y_i given the value x_i of the scalar covariate x and error terms are independent and follow the normal distribution $N(0, 1)$. To pursue estimation, $r(\cdot)$ is replaced by a parametric regression spline model

$$r(x; \boldsymbol{\beta}, \mathbf{b}) = \beta_1 + \beta_2 x + \dots + \beta_p x^{p-1} + \sum_{j=1}^m b_j z_j(x). \quad (2.2.13)$$

The first p terms are a $(p - 1)$ th order polynomial of x , the covariates $z_1(x), \dots, z_m(x)$ are elements of a smoothing basis, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ and $\mathbf{b} = (b_1, \dots, b_m)'$ are unknown parameters. Then (2.2.13) can be written as

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \mathbf{z}'_i \mathbf{b} + \sigma \varepsilon_i,$$

where $\mathbf{x}_i = (1, x_i, \dots, x_i^{p-1})'$ and $\mathbf{z}_i = (z_1(x_i), \dots, z_m(x_i))'$. Typically \mathbf{x}_i is low-dimensional and \mathbf{z}_i is a high-dimensional basis linearly independent of \mathbf{x}_i . A convenient choice is to use the truncated power basis of degree $p-1$. Then the i th row of \mathbf{Z} is $\mathbf{z}_i = ((x_i - \kappa_1)_+^{p-1}, \dots, (x_i - \kappa_m)_+^{p-1})$ with x_+ as positive part, so that for any number x , x_+ is x if x is positive and is equal to 0 otherwise. The knots $\kappa_1, \dots, \kappa_m$ are fixed values covering the range of x_1, \dots, x_n .

The model (2.2.13) is represented as a linear mixed model

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}, & \mathbf{b} &\sim \mathbf{N}(\mathbf{0}, \phi^2 \mathbf{I}_m), \\ \boldsymbol{\varepsilon} &\sim \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n), & \text{Cov}(\mathbf{b}, \boldsymbol{\varepsilon}) &= \mathbf{0}, \end{aligned}$$

where \mathbf{X} and \mathbf{Z} are known $n \times p$ and $n \times m$ matrices, respectively, \mathbf{b} is the $m \times 1$ vector of random effects that occur in the $n \times 1$ data vector \mathbf{y} and $\boldsymbol{\beta}$ is the $p \times 1$ vector of unknown fixed effects parameters. Compared with the ordinary linear regression model, the difference is $\mathbf{Z}\mathbf{b}$, which may take various forms, thus creating a rich class of models. Then under these conditions we have

$$\mathbf{y} \sim \mathbf{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{V})$$

and

$$\mathbf{y}|\mathbf{b} \sim \mathbf{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \sigma^2 \mathbf{I}_n), \quad (2.2.14)$$

where $\mathbf{V} = \frac{1}{\alpha} \mathbf{Z}\mathbf{Z}' + \mathbf{I}_n$ for $\alpha = \sigma^2/\phi^2 > 0$. The parameter α is the ratio between the variance of the error variables ε_i , $1 \leq i \leq n$ and the variance of the random effects b_j , $1 \leq j \leq m$. The set of possible values for α is $[0, \infty]$.

The amount of smoothing is controlled by α , which is here referred to as a smoothing parameter. In addition to the value of α , the degree of the regression spline and the number and location of knots must be specified. In the LMM (2.2.14) the interest is either in the fixed effects parameter $\boldsymbol{\beta}$, or also in the associated random effects \mathbf{b} . We derive the ML estimates of $\boldsymbol{\beta}$ and \mathbf{b} in [6].

2.2.2 MLE in logistic regression

The covariates (comorbidities, age, sex and type of hip fracture) in our model (1.1.3) are such that typically more than one patient has a fixed \mathbf{x}_i value, i.e. the setting i of k explanatory variables, and hence the number of different settings l is less than n . Therefore it is sufficient to record the number of observations n_i and the number of deaths v_i corresponding to the different settings $i = 1, \dots, l$. We let v_i refer to this death count rather than to an individual binary response, and then $\mathbf{v} = (v_1, \dots, v_l)'$ is an l -dimensional vector of independent binomials. Then the likelihood (1.1.3) takes the form

$$L(\boldsymbol{\beta}; \mathbf{x}_1, \dots, \mathbf{x}_n) = \prod_{i=1}^l \pi(\mathbf{x}_i; \boldsymbol{\beta})^{v_i} [1 - \pi(\mathbf{x}_i; \boldsymbol{\beta})]^{n_i - v_i},$$

where $v_i = \sum_{t: x_t = x_i} y_t$ is the number of deaths among the patients with the setting \mathbf{x}_i , $i = 1, \dots, l$.

Let Γ be the set of all $1 \times k$ vectors of the form $\gamma = (\gamma_1, \dots, \gamma_k)$, where $\gamma_j = 0$ or 1 for $j = 1, \dots, k$. There are 2^k such vectors in Γ . A variable selection procedure is then equivalent to first selecting $\gamma \in \Gamma$. If $\gamma_j = 1$, the variable x_j , $1 \leq j \leq k$ is selected and the corresponding β_j is estimated, otherwise $\gamma_j = 0$ and $\beta_j = 0$, i.e. x_i is not selected. Let $\boldsymbol{\beta}_\gamma = \text{diag}[\gamma]\boldsymbol{\beta}$, where $\text{diag}[\gamma]$ is the $k \times k$ diagonal matrix with diagonal elements γ and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)'$ is the k -dimensional parameter vector. We suppose that the independent binomials $\{v_1, \dots, v_l\}$ with $\mathbf{E}(v_i) = n_i \pi(\mathbf{x}_i; \gamma)$ are related to the covariates by the logistic regression model

$$\pi(\mathbf{x}_i; \boldsymbol{\beta}_\gamma) = \exp(\boldsymbol{\beta}_\gamma' \mathbf{x}_i) / [1 + \exp(\boldsymbol{\beta}_\gamma' \mathbf{x}_i)], \quad (2.2.15)$$

where $n = n_1 + \dots + n_l$.

The likelihood function $L(\boldsymbol{\beta}_\gamma | \mathbf{v}; \mathbf{X})$, under the model γ , is proportional to the product of l binomial functions

$$\begin{aligned} & \prod_{i=1}^l \pi(\mathbf{x}_i; \gamma)^{v_i} [1 - \pi(\mathbf{x}_i; \gamma)]^{n_i - v_i} \\ &= \left\{ \exp \left[\sum_{i=1}^l v_i \log \frac{\pi(\mathbf{x}_i; \gamma)}{1 - \pi(\mathbf{x}_i; \gamma)} \right] \right\} \left\{ \prod_{i=1}^l [1 - \pi(\mathbf{x}_i; \gamma)]^{n_i} \right\}. \end{aligned} \quad (2.2.16)$$

For model (2.2.15), the i th logit is $\boldsymbol{\beta}_\gamma' \mathbf{x}_i$, so the exponential term in (2.2.16) equals $\exp(\mathbf{v}' \mathbf{X} \boldsymbol{\beta}_\gamma)$, where the $l \times k$ regressor matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_l)'$ denotes the l different settings of covariates. Since

$$[1 - \pi(\mathbf{x}_i; \gamma)] = [1 + \exp(\boldsymbol{\beta}_\gamma' \mathbf{x}_i)]^{-1},$$

the log likelihood equals

$$l(\boldsymbol{\beta}_\gamma | \mathbf{v}; \mathbf{X}) = \mathbf{v}' \mathbf{X} \boldsymbol{\beta}_\gamma - \sum_{i=1}^l \log[1 + \exp(\boldsymbol{\beta}_\gamma' \mathbf{x}_i)].$$

The maximum likelihood estimate $\hat{\boldsymbol{\beta}}_\gamma$ of $\boldsymbol{\beta}$ for γ is obtained by solving the likelihood equations which result from setting $\partial l(\boldsymbol{\beta}_\gamma | \mathbf{v}; \mathbf{X}) / \partial \boldsymbol{\beta}_\gamma = 0$, and they may be written in the form

$$\mathbf{X}' \mathbf{v} = \mathbf{X}' \hat{\boldsymbol{\mu}}$$

where $\hat{\mu}_i = n_i \pi(\mathbf{x}_i; \hat{\boldsymbol{\beta}}_\gamma)$, $i = 1, \dots, l$, is the i th element of the $l \times 1$ vector $\hat{\boldsymbol{\mu}}$. Rissanen's (1996) NML distribution is obtained by normalizing the maximum likelihoods

$$L[\hat{\boldsymbol{\beta}}_\gamma(\mathbf{v}) | \mathbf{v}; \mathbf{X}] \quad (2.2.17)$$

over the data. The NML distribution is used in [7] as a basic tool in statistical modeling.

3

Model selection with MDL

There are different approaches to model selection, depending on the aims and uses associated with the selected model. Most MS methods are formulated in terms of an information criterion that uses data to give each model a score such that we have a ranked list of competing models. In this chapter we deal with the MDL approach to modeling and we illustrate it with four specific modeling problems which we have studied in articles [1], [2], [5], [6] and [7].

3.1 Introduction to modeling with MDL

This section is a short introduction to the MDL approach which is based on normalized maximum likelihood.

3.1.1 Prefix Codes

Let Y be a discrete random variable with the probability mass function $f(y) = P(Y = y)$ and the support $\mathcal{Y} \subset \mathbb{R}$ such that $f(y) > 0$ for $y \in \mathcal{Y}$. A binary code C is a mapping from \mathcal{Y} to a set of codewords which are finite-length strings of bits. Let $C(y)$ denote the codeword corresponding to y and l denotes the codelength function associated with this code C so that $l(y)$ is the code length of y . The map C is required to be one-to-one, and concatenations of codewords are also required to be in one-to-one correspondence with sequences of symbols from \mathcal{Y} . This requirement of unique decodability is accomplished in particular by arranging the codewords to satisfy the property that no codeword is a prefix for a codeword of another $y \in \mathcal{Y}$. Therefore C is assumed to be a prefix code (Cover and Thomas 1991, Barron et al. 1998), no codeword is a prefix of any other codeword.

3.1.2 Codelengths and probabilities

Already Shannon posed the problem: What codelengths achieve the minimum expected value $\mathbf{E}_f[l(Y)]$? The solution to the problem is to take $l^*(y) = \log[1/f(y)]$ if we ignore the integer codelength constraint. The solution $\log[1/f(y)]$ is called the ideal codelength or the Shannon codelength (Cover and Thomas 1991). With any other choice of a probability mass function, say q , the excess codelength

$$l(y) - l^*(y) = \log[1/q(y)] - \log[1/f(y)] = \log \frac{f(y)}{q(y)}$$

has positive expected value $\mathbf{E}_f \log[f(Y)/q(Y)]$, the Kullback-Leibler distance, which is zero only if $f = q$. So, $l^*(y)$ is the optimal codelength of y . In general, there is a correspondence between codelengths and a probability distribution on \mathcal{Y} . An integer-valued function l corresponds to the code length of a binary prefix code if and only if it satisfies Kraft's inequality,

$$\sum_{y \in \mathcal{Y}} 2^{-l(y)} \leq 1.$$

Therefore, for a given prefix code C on \mathcal{Y} with length function l , we can define a distribution on \mathcal{Y} as

$$q(y) = 2^{-l(y)} / K \quad \text{for } y \in \mathcal{Y},$$

where K denotes the sum on the left side of the Kraft's inequality. Conversely, for any distribution q on \mathcal{Y} and any $y \in \mathcal{Y}$, we can find a prefix code with length function $l(y) = \lceil \log(1/q(y)) \rceil$, the smallest integer greater than or equal to $\log(1/q(y))$. From this point of view, a codelength is just another way to express a probability distribution (Cover and Thomas 1991). We are not concerned with actual codings, but we are only concerned with code length functions. A short codelength corresponds to a high probability and vice versa.

3.1.3 Normalized maximum likelihood

An extension of Shannon's theory can be obtained if instead of one fixed distribution f we suppose a model class, a parametric family of probability mass functions

$$\mathcal{F} = \{f(y; \theta) : \theta \in \Theta \subset \mathbb{R}^k\},$$

which have the ideal codelengths $\log[1/f(y; \theta)]$ for $y \in \mathcal{Y}$. After observing y , the shortest codelength for y is $\log[1/f(y; \hat{\theta})]$, where $\hat{\theta} = \hat{\theta}(y)$ is the ML estimate of θ . For a given y , $f(y; \theta)$ is the likelihood function of θ and $f(y; \hat{\theta}) = \max_{\theta} f(y; \theta)$. Note, however, that $f(y; \hat{\theta}(y))$ does not define a distribution for $y \in \mathcal{Y}$. If we use a distribution q , the excess code length is

$$\log[1/q(y)] - \log[1/f(y; \hat{\theta})] = \log \frac{f(y; \hat{\theta})}{q(y)}. \quad (3.1.1)$$

Shtarkov (1987) posed the problem of choosing q to minimize the worst case value $\max_{y \in \mathcal{Y}} \log[f(y; \hat{\theta})/q(y)]$ of (3.1.1), and he found the normalized maximum likelihood (NML)

$$\hat{f}(y) = \frac{f(y; \hat{\theta}(y))}{C(\mathcal{F})} \quad (3.1.2)$$

as the unique solution, where $C(\mathcal{F}) = \sum_{y \in \mathcal{Y}} f(y; \hat{\theta}(y))$.

This NML distribution has an important role in the MDL theory. Now $\hat{f}(y)$ does not depend on any unknown parameter, and hence the codelength corresponding to it can be computed. The codelength corresponding to the NML distribution (3.1.2) is

$$\log[1/\hat{f}(y)] = -\log f(y; \hat{\theta}) + \log C(\mathcal{F}). \quad (3.1.3)$$

This optimal codelength $\log[1/\hat{f}(y)]$ associated with the NML distribution is called the *stochastic complexity* (SC) of data relative to the model class \mathcal{F} , and clearly it depends on the model class \mathcal{F} . The additional codelength $\log C(\mathcal{F})$ due to the unknown parameter, is called the *parametric complexity*. In the case of continuous random variables we may replace the sum in $C(\mathcal{F})$ by the corresponding integral. Then all results remain virtually unchanged when probability mass functions are replaced by density functions and sums by integrals. Rissanen (1996) introduced the concepts NML and SC as tools of statistical inference and model selection. Good introductions to these ideas are Barron et al. (1998) and Hansen and Yu (2001), for example.

One may raise the question of how to initially decide the specific family of models \mathcal{F} . A direct quotation from Rissanen's book (2007, p. 101) gives a good answer. "In conclusion, it is important to realize that the MDL principle has nothing to say about how to select the suggested family of model classes. In fact, this is a problem that cannot be adequately formalized. In practice their selection is based on human judgement and prior knowledge of the kinds of models that have been used in the past, perhaps by other researchers."

3.2 Variable selection in linear regression

The problem of variable selection is one of the most pervasive problems in statistical applications. One wants to model the relationship between y and a subset of potential explanatory variables x_1, \dots, x_p , but there is uncertainty about which subset to use. Letting γ index the subsets of the columns $\mathbf{x}_1, \dots, \mathbf{x}_p$ of \mathbf{X} in (1.1.2), the problem is to select and fit a model of the form

$$\{y, \mathbf{X}_\gamma \beta_\gamma, \sigma^2 \mathbf{I}_n\}, \quad (3.2.4)$$

where \mathbf{X}_γ is an $n \times q_\gamma$ matrix whose columns correspond to the γ th subset, q_γ is the size of the γ th subset and β_γ is a $q_\gamma \times 1$ vector of regression

coefficients. Let $f(\mathbf{y}; \boldsymbol{\theta}_\gamma)$ denote the density function of \mathbf{y} under the model (3.2.4), where $\boldsymbol{\theta}_\gamma = (\boldsymbol{\beta}'_\gamma, \sigma^2)'$. We compute for each γ in (3.2.4) its SC $-\log \hat{f}(\mathbf{y}; \gamma)$ and then according to the MDL principle we determine the model $\gamma = \hat{\gamma}$ that minimizes the SC:

$$-\log \hat{f}(\mathbf{y}; \hat{\gamma}) = \min_{\gamma} \{-\log \hat{f}(\mathbf{y}; \gamma) : \gamma \in \Gamma\},$$

where $\Gamma = \{1, 2, \dots, 2^p\}$.

Rissanen (1996) introduced an MDL criterion based on the NML coding scheme of Shtarkov (1987) and developed it as a tool for statistical modeling and inference. However, it turned out that the parametric complexity $C(\mathcal{F})$ for some families of distributions, e.g. for the important normal distribution, was not finite and hence the definition of the NML distribution (3.1.2) fails in these cases. For the Gaussian density the parametric complexity $C(\mathcal{F})$ is the integral of the maximized likelihood $f(\mathbf{y}; \hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y}))$ over \mathbb{R}^n which is not finite for (1.1.2). Then Rissanen (1996) and Barron et al. (1998) suggested to restrict the integration domain as follows

$$\mathcal{V}(s, R) = \{\mathbf{y} : \hat{\boldsymbol{\beta}}(\mathbf{y})' \mathbf{X}' \mathbf{X} \hat{\boldsymbol{\beta}}(\mathbf{y}) \leq nR, \hat{\sigma}^2(\mathbf{y}) \geq s\}, \quad (3.2.5)$$

where s and R are given positive constants, and $\hat{\sigma}^2$ is the ML estimate of σ^2 . Then we have finite parametric complexity $C(s, R)$, i.e. the normalizing constant of the NML distribution, which depends on the hyperparameters s and R . One option is to use $C(\hat{s}, \hat{R})$ obtained by replacing s and R with their ML estimates \hat{s} and \hat{R} as suggested by Barron et al. (1998) and Hansen and Yu (2001). The resulting "NML function"

$$\hat{f}(\mathbf{y}; \hat{s}, \hat{R}) = f(\mathbf{y}; \hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y})) / C(\hat{s}, \hat{R}) \quad (3.2.6)$$

is not, however, a density function. Therefore Rissanen (2000) applied normalization on the function (3.2.6). Essentially, the idea is to treat the hyperparameters s and R as the parameters σ^2 and β in the first phase. He found that this second normalization makes the effect of the hyperparameters on the resulting code length additive and hence can be ignored for model selection.

However, replacing the first restriction in (3.2.5) with a general ellipsoidal constraint

$$\hat{\boldsymbol{\beta}}(\mathbf{y})' \mathbf{Q} \hat{\boldsymbol{\beta}}(\mathbf{y}) \leq nR \quad (3.2.7)$$

affects the criterion (3.1.3) in an essential manner. The matrix \mathbf{Q} in (3.2.7) is positive definite. This will open a new way to extend the scope and applications of the MDL principle. Usually MS criteria work well when two models with different number of estimated parameters are compared, but they may not discriminate well between models with the same number of estimated parameters. If the regressors are near collinear, the determinant $|\mathbf{X}'_\gamma \mathbf{X}_\gamma| \approx 0$. We may want a criterion which detects multicollinearity, and arranges the models of a fixed size k according to the severity of multicollinearity. So, the number of estimated parameters and model fit would

not be the only features when considering the quality of a criterion. Originally, the wish to find such criteria motivated us to study the MDL criterion from the point of view of data restrictions. So, this approach serves as a technique to tailor an MS criterion that detects a certain specific feature of a model. We also extended our study beyond ellisoidal restrictions, and we believe that there are promising prospects in this field. Further research on the effects and interpretations of the constraints in various applications is needed.

3.3 Spline smoothing

The amount of smoothing in the spline model (2.2.13) is controlled by α , which is here referred to as a smoothing parameter. The parameter $\alpha = \sigma^2/\phi^2 > 0$ is defined as the ratio between the variance of the error variables and the variance of the random effects in the LMM (2.2.14). The set of possible values for α is $[0, \infty]$. We adopt the procedure where the knots are located at "equally spaced" sample quantiles of x_1, \dots, x_n . Thus the k th knot is the j th order statistic of $x_{(1)}, \dots, x_{(n)}$ where j is $nk/(m+1)$ rounded to the nearest integer. As soon as the degree of the regression spline is specified, one has to fix the number of knots. It is often recommended to choose the basis in a "generous" manner such that there are enough knots to fit features in the data (see e.g. Ruppert et al. 2002).

In smoothing we control three modeling parameters: the degree of the regression spline $p-1$, the number of knots m and the smoothing parameter α . A model $\gamma = (p, m, \alpha)$ is specified by the triple where the values for the modeling parameters p, m and α should be determined in an optimal way. The choice of α has a profound influence on the fit. In fact, it was shown in [6] that α can be chosen to give any one of a spectrum of fits between the unconstrained regression spline fit and the least-squares polynomial fit.

The MDL model selection criterion $MDL(\gamma)$ for spline smoothing is derived in [6] by using a technique similar to that used in Gaussian linear regression. A model estimate $\hat{\gamma}$ is obtained by minimizing the $MDL(\gamma)$ selection criterion with respect to model $\gamma = (p, m, \alpha)$, that is, with respect to parameters p, m and α , using numerical optimization routines.

3.4 Sequential NML in logistic regression

We consider now the NML criterion in the case of logistic regression. The NML function for (2.2.17) may now be written as

$$\hat{P}(\mathbf{v}|\gamma) = L[\hat{\beta}_\gamma(\mathbf{v})|\mathbf{v}; \mathbf{X}]/C(\gamma),$$

where $L[\hat{\beta}_\gamma(\mathbf{v})|\mathbf{v}; \mathbf{X}]$ is the maximum of the likelihood function and

$$C(\gamma) = \sum_{\mathbf{v} \in \Omega} L[\hat{\beta}_\gamma(\mathbf{v})|\mathbf{v}; \mathbf{X}] \quad (3.4.8)$$

is the normalizing constant. In (3.4.8) Ω denotes the sample space and the sum runs over all different count vectors (v_1, \dots, v_l) such that $0 \leq v_1 + \dots + v_l \leq n$ and $v_i \geq 0$, $i = 1, \dots, l$. The notation $\hat{\beta}_\gamma(\mathbf{v})$ emphasizes the obvious fact that the ML estimate $\hat{\beta}_\gamma$ is a function of \mathbf{v} .

The summation in (3.4.8) is over all count vectors $\mathbf{v}_\gamma = (v_1, \dots, v_{l_\gamma})$ such that $0 \leq v_1 + \dots + v_{l_\gamma} \leq n$ and $0 \leq v_i \leq n_i$, $i = 1, \dots, l_\gamma$, where $n = 28797$ and l_γ denotes the number of different settings of covariate values in the data under the model γ . Since the ML estimate $\hat{\beta}_\gamma(\mathbf{v}_\gamma)$ and $L[\hat{\beta}_\gamma(\mathbf{v}_\gamma)|\mathbf{v}_\gamma; \mathbf{X}]$ has to be computed over all possible count vectors, it is obvious that the computation of the normalization constant just for one model γ is heavy, not to mention the situation where we wish to compare all competing models. Therefore we introduce a new *MDL* based model selection criterion following the idea of sequentially normalized maximum likelihood (sNML) that was proposed by Rissanen and Roos (2007).

Roos and Rissanen (2008) presented the sequentially normalized maximum likelihood (sNML) function. Let $X^n = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ denote the matrix of regressors and $\mathbf{y}^n = (y_1, \dots, y_n)$ a sequence of the binary outcome variables. Note that here \mathbf{x}_i denotes the regressor vector of the i th patient and X^n may contain identical regressor vectors unlike \mathbf{X} in (3.4.8). In the logistic regression case, the sNML function may be written as

$$\hat{P}(\mathbf{y}^n | X^n) = \hat{P}(\mathbf{y}^m | X^m) \prod_{t=m+1}^n \hat{P}(y_t | \mathbf{y}^{t-1} X^t), \quad (3.4.9)$$

where $\hat{P}(\mathbf{y}^n | X^n)$ is the estimated probability to observe the string \mathbf{y}^n having observed X^n .

The last term from (3.4.9) is the NML function for y_t

$$\hat{P}(y_t | \mathbf{y}^{t-1} X^t) = \frac{P(y_t | \mathbf{y}^{t-1}, X^t, \hat{\beta}(\mathbf{y}^t))}{K(\mathbf{y}^{t-1})}, \quad (3.4.10)$$

where

$$K(\mathbf{y}^{t-1}) = P(y_t = 0 | \mathbf{y}^{t-1}, X^t, \hat{\beta}_0) + P(y_t = 1 | \mathbf{y}^{t-1}, X^t, \hat{\beta}_1)$$

is the normalizing constant. Here $\hat{\beta}_i$ denotes the ML estimates of β from the binary outcome vectors $(\mathbf{y}^{t-1}, 0)$ and $(\mathbf{y}^{t-1}, 1)$, respectively. In (3.4.10) we normalize only over the last observation which drastically simplifies the computation of the normalizing constant compared to the standard NML.

Because the observations are independent, the negative logarithm of the

sNML function (3.4.9) takes the form

$$\begin{aligned}
-\log \hat{P}(\mathbf{y}^n, X^n) &= -\log \hat{P}(\mathbf{y}^m | X^m) - \sum_{t=m+1}^n \hat{P}(y_t | \mathbf{y}^{t-1}, X^t) \\
&= -\log \hat{P}(\mathbf{y}^m | X^m) - \sum_{t=m+1}^n \log P(y_t | \mathbf{y}^{t-1}, X^t, \hat{\beta}(\mathbf{y}^t)) \\
&\quad + \sum_{t=m+1}^n \log K(\mathbf{y}^{t-1}).
\end{aligned} \tag{3.4.11}$$

The computational load of $\sum_{t=m+1}^n \log K(\mathbf{y}^{t-1})$ in (3.4.11) is trivial compared to the load of $\log [C(\gamma)]$ in (3.4.8). Hence the sNML criterion is applicable also to wide models with large amounts of data like our risk-adjustment model for hip fracture mortality in Finland.

3.5 Weight selection in nonparametric regression

In article [5] we assume that the data follow a classical nonparametric regression model

$$y_i = \mu(\mathbf{x}_i) + \sigma \varepsilon_i, \quad 1 = 1, \dots, n, \tag{3.5.12}$$

where $\varepsilon_1, \dots, \varepsilon_n$ are independent random variables. We have observations $(y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)$, where y_1, \dots, y_n are real valued and $\mathbf{x}_i = (x_{i1}, \dots, x_{ik_M})'$ is a $k_M \times 1$ vector such that for each $1 \leq i \leq n$,

$$\mathbf{E}(\varepsilon_i | \mathbf{x}_i) = 0 \quad \text{and} \quad \mathbf{E}(\varepsilon_i^2 | \mathbf{x}_i) = 1,$$

and σ is the scale parameter of the additive error $\sigma \varepsilon$. Here we assume that μ is in the space of square integrable functions L_2 whose elements admit representations as infinite dimensional linear models for which

$$\mu(\mathbf{x}) = \sum_{j=1}^{\infty} \beta_j \varphi_j(\mathbf{x}) \tag{3.5.13}$$

for some set of known functions $\{\varphi_1, \varphi_2, \dots\}$ and real valued coefficients β_1, β_2, \dots .

The practical significance of (3.5.13) is that any $\mu \in L_2$ may be well approximated by a finite number of m leading terms in (3.5.13):

$$\mu_m(\mathbf{x}) = \sum_{j=1}^m \beta_j \varphi_j(\mathbf{x}),$$

and we denote generally $x_{ij} = \varphi_j(\mathbf{x}_i)$. To obtain an estimate of μ one may employ an approximating linear model

$$y_i = \sum_{j \in \mathcal{M}_m} x_{ij} \beta_j + b_i + \sigma \varepsilon_i, \quad i = 1, 2, \dots, n,$$

where $\mathcal{M}_m = \{1, 2, \dots, k_m\}$ with $k_m \leq n$, the approximation error $b_i = \sum_{j=k_m+1}^{\infty} \beta_j x_{ij}$ and the random errors $\varepsilon_1, \dots, \varepsilon_n$ are like in (3.5.12). The set of approximating models $\{\mathcal{M}_1, \dots, \mathcal{M}_M\}$ is such that $\mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq \dots \mathcal{M}_M$.

The response data \mathbf{y} are modeled with the normal density functions

$$f(\mathbf{y}; \boldsymbol{\beta}_m, \sigma_m^2) = \frac{1}{(2\pi\sigma_m^2)^{n/2}} \exp\left(-\frac{1}{2\sigma_m^2} \|\mathbf{y} - \boldsymbol{\mu}_m\|^2\right),$$

where $\boldsymbol{\mu}_m = \mathbf{X}_m \boldsymbol{\beta}_m$, \mathbf{X}_m is the $n \times k_m$ matrix with ij element x_{ij} and $\boldsymbol{\beta}_m = (\beta_1, \dots, \beta_{k_m})'$ is the $k_m \times 1$ vector of unknown parameters. The matrix \mathbf{X}_m is of full column rank and $\|\cdot\|$ denotes the Euclidean norm.

To deal with model uncertainty, we average over the estimates across the set of approximating models. We consider estimators $\hat{\boldsymbol{\mu}}_w$ of $\boldsymbol{\mu}_w$ which are convex combinations of the ML estimators $\hat{\boldsymbol{\mu}}_m$, $m = 1, \dots, M$:

$$\hat{\boldsymbol{\mu}}_w = \sum_{m=1}^M w_m \hat{\boldsymbol{\mu}}_m,$$

where $\sum_{m=1}^M w_m = 1$, $w_m \geq 0$, $m = 1, \dots, M$.

In MDL weighting we consider a mixture density

$$\sum_{m=1}^M w_m \hat{f}(\mathbf{y}; m), \quad \text{with } w_m \geq 0, \quad \sum_{m=1}^M w_m = 1,$$

where $\hat{f}(\mathbf{y}; m)$, $m = 1, \dots, M$, are NML densities for Gaussian linear regressions. If we select the model $m = \hat{m}$ and encode the data using the selected model \hat{m} , then $\log \{1/[w_{\hat{m}} \hat{f}(\mathbf{y}; \hat{m})]\}$ is the code length for the data. On the other hand, the mixture yields the code length $\log \{1/[\sum_{m=1}^M w_m \hat{f}(\mathbf{y}; m)]\}$ which is always shorter if $w_{\hat{m}} \neq 1$. Therefore, it seems advantageous to encode with a mixture (see also Liang and Barron 2005). However, the problem of finding the weights vector still remains.

Given the data \mathbf{y} , $\hat{f}(\mathbf{y}; m)$ can be interpreted as the likelihood of the model \mathcal{M}_m . This leads to the NML distribution for the models:

$$\hat{p}(m; \mathbf{y}) = \frac{\hat{f}(\mathbf{y}; m)}{\sum_{i=1}^M \hat{f}(\mathbf{y}; i)} = \frac{\exp(-MDL_m/2)}{\sum_{i=1}^M \exp(-MDL_i/2)}, \quad (3.5.14)$$

where $MDL_m = -2 \log \hat{f}(\mathbf{y}; m)$ denotes the value of MDL model selection criterion for the model \mathcal{M}_m . Thus the MDL distribution (3.5.14) may be used to define the empirically selected weights $\hat{w}_m = \hat{p}(m; \mathbf{y})$ for models \mathcal{M}_m , $m = 1, \dots, M$.

4

Applications to hip fracture data

Hip fractures are an important cause leading to raised morbidity and mortality among the elderly population (Liporace *et al.* 2005). They do not only affect the patient himself, but also cause remarkable costs to the society (Hannan *et al.* 2001). Between 1999 – 2007 the mean cost of a hip fracture patient during the first year after the fracture has gone up from 18000 to almost 20000 euros (Sund *et al.* 2011). For a first time hip fracture patient the costs during the first year after the fracture were estimated to be around 14410 euros in Finland (Nurmi *et al.* 2003). About 7000 hip fractures occur per year in Finland (Sund 2006). The majority of fractures occur to persons over 50 years of age. Among patients of 50 years and older hip fracture is usually caused by a low energy trauma such as falling from standing height or lower (Zuckerman 1996). Among younger patients the fracture is usually caused by a high-energy trauma such as traffic accident or falling from a height (Robinson *et al.* 1995).

4.1 The PERFECT project

The PERFECT (PERFormance, Effectiveness and Cost of Treatment episodes) project was started as a co-operative project between hospital districts, the Social Insurance Institute and the National Research and Development Centre for Welfare and Health (STAKES, nowadays THL) in the year 2004. The aim of the project is to measure the effectiveness and costs of seven major diseases using existing linkable information available from registers. The diseases studied in the project are stroke, myocardial infarction, breast cancer, schizophrenia, very preterm infants, hip fracture, and total hip arthroplasty. The hip fracture data from the years 1999 – 2005 used in the articles [3], [4] and [7] of this thesis belongs to the PERFECT project and was kindly provided to the author with permission from STAKES and THL by Professor Unto Häkkinen (THL).

The production of the dataset has required several years of work from a multidisciplinary team of experts. Also the author was involved in this work

during 2005–2006. A more detailed description on the choices made during the production of the dataset (e.g. definition of hip fracture, inclusion and exclusion criteria) can be found in Sund et al. (2011). Only patients of 50 years or older are in the hip fracture dataset of the PERFECT project. This is because the trauma mechanism is practically always different for persons under 50 years of age (Sund et al. 2011). Another important exclusion criterion was institutionalization. Institutionalized long-term-care patients were excluded from the dataset because their expected outcomes and use of resources (as well as appropriate performance indicators) differ significantly from home-dwelling patients (Sund et al. 2008).

4.2 Medical care costs of hip fracture treatments

A hip fracture is usually very painful and needs hospital treatment. Therefore virtually all hip fractures are recorded in national registries. As a patient is discharged from the hospital, information such as diagnosis, operation and length of stay are recorded. Finland has a long tradition of collecting data on social and health services and the Finnish national registries are exceptional world wide (Gissler 1999, Gissler and Haukka 2004). A review on the quality of the Finnish Hospital Discharge Register may be found in Sund (2012). With proper risk-adjustment and data handling Finnish register data may be used for system and producer-level performance measurement (Peltola *et al.* 2011). When handled correctly, the data from national registers has also been found suitable for the performance assessment of hip fracture treatment (Sund et al. 2007).

The costs for hip fracture patients were estimated by using prices based on diagnosis related grouping (DRG) and the number of bed days during the treatment episodes. For the estimation of costs for hip fracture operations more detailed cost data from Helsinki and Uusimaa district were used. The costs were inflation adjusted to the level of the year 2005. More information on the estimation of the treatment costs used in this thesis can be found in Peltola et al. (2011). In [3] we compare the hip fracture treatment costs between 7 hospital districts and contrast WALS estimation with backwards elimination technique. In [4] we utilize the hip fracture data in simulation experiments to compare the performance of various penalized LS estimators, including the restricted LS estimator, within a realistic set-up.

4.3 Risk-adjustment model for hip fracture mortality

If we wish to compare hospitals or hospital districts with respect to a performance indicator (e.g. mortality), risk-adjustment is often desirable to account for possible differences in patient case mix (Iezzoni 2003). Vari-

ables such as sex and age on admission are usually quite straightforward to include in the model, but we may want to adjust also for other variables. The patient may for example have diseases or disorders on the event of the hip fracture that have an impact on the outcome of the treatment. We call these diseases or disorders comorbidities. In some cases equally important as finding indication for a comorbidity is finding the timepoint when that comorbidity occurred for the first or last time. In article [7] we wish to find an answer to the problem: How much medical history do we need to effectively adjust for congestive heart failure, cancer and diabetes? The criteria for selecting comorbidities are not only statistical but for example a high prevalence of comorbidity and the effect of comorbidity on the treatment of hip fracture are important aspects.

The main objective in the treatment of hip fracture is to help the patient regain his/her pre-fracture health status and level of functional ability. Because a successful treatment should make it possible that patients are able to continue life in the same fashion as before the fracture, death is obviously a very unsuccessful outcome.

Hip fracture itself does not necessarily lead to death, but especially for patients in a lowered physical condition before the hip fracture, it may trigger a process that ultimately leads to death (Heithoff and Lohr 1990). If the hip fracture triggers the dying process, we may assume that short-term mortality is in fact an indicator that the patient's health status before the hip fracture was already substantially lowered. Hip fracture serves as a tracer condition in the performance assessment of health systems because it provides a large group of vulnerable patients to study how well health and social services are integrated in the provision of acute care and rehabilitation (SIGN 2002).

In [7] we have chosen two mortality indicators, 90 days mortality and one year mortality. Every patient has at least one year follow up so no censoring was present. Another reason for these indicators is that they are widely used and an interpretation can be found for both of them. Mortality is also well defined and an easily observable indicator in the sense that there is typically no argument if a patient is dead or not. The 90 days mortality reflects the risk connected to hip fracture treatment and one year mortality reflects more the overall condition of a patient than risk of death caused directly by the shock effect of the hip fracture event. We could use also continuous responses, but that would make the interpretation of results a bit more laborous and require a different method. Sund (2008) demonstrates how complicated responses can be reduced to simpler summarizations and further to more traditional indicators in the case of hip fracture data.

In article [7] we wish to find factors that explain the mortality following hip fracture, measured as a binary variable, in order to obtain a set of covariates which profile a patient's medical condition at the time of the hip fracture. Our interest is in three comorbidities that a patient has had before the hip fracture and which may have effect on the outcome of the treatment. The special focus in our study is to examine how far we have to

follow the patients medical history. Various lengths of the follow-up period are modeled in order to find the shortest period to effectively adjust for each comorbidity.

In the dataset (backwards) hospitalization history was available up to 10 years before the fracture. This information was complemented with data obtained from the register maintained by the Social Insurance Institution of Finland. From this second register, information on drug reimbursements was obtained. The mortality was followed using the Causes of Death register of Statistics Finland.

There were two ways to get an indication for a comorbidity from our data. Firstly, we have data on a patient's all hospitalization periods preceding the hip fracture until a certain (historical) time point. Now if the patient has been hospitalized because of a certain comorbidity between this time point and the hip fracture, we get indication that the patient has had that comorbidity. The second way to get indication of a comorbidity comes through information on drug reimbursements. Now we have to check if a patient has received the right for drug reimbursements for that comorbidity and that it was still valid when the hip fracture occurred. This means that if a patient has had the right for drug reimbursements when the hip fracture occurred, then the patient will have indication for that comorbidity for all time periods.

Our analysis in [7] is meant to be a preliminary analysis in constructing the risk-adjustment model which we use to compare hospitals or hospital districts. In practice we may have to go through over a hundred comorbidities. Therefore we wanted to keep the setting very simple and use the same historical timepoints for each comorbidity. Medical history could be used also in a continuous manner in logistic regression. Another simplification that we made was omitting interactions. This was because interactions rarely became significant. Hietala (2009) analysed the same data with logistic regression and detected certain statistically significant interactions, for example between age and cancer, and between cancer and gender. However, interactions did not have a pivotal role in the statistically best models he found. Further, an acceptable medical interpretations of interactions should be available. Also in PERFECT the practice was to omit interactions to keep the models as simple as possible. Simplicity is sought because of the high number of indicators developed in the project. Another reason is that the indicators are produced annually, so also the indicator specific methods ought to be updated annually (Peltola et al. 2011). We didn't want to deviate from these practices so that our results could possibly be used in PERFECT.

The setting in [7] is actually quite challenging from the model selection point of view, since the number of the occurrences of a disease does not increase much when the length of inspection period increases. If we change our view for example from 180 days to one year before the fracture, the increase in the number of occurrences is typically small, especially when compared to the size of the whole data. Therefore it may be difficult to

distinguish between models that use different time period variables. If we look further back in history, more occurrences appear, but the effect of these occurrences on the dependent variable may become weaker. We assume that this time dependence may not be the same for all comorbidities.

5

Summary of publications and author's contribution

The thesis consists of seven research articles, five of them are published and two are accepted for publication. In this chapter we present a summary of each of the articles and explain the author's contribution to them.

5.1 Summary of publications

The normalized maximum likelihood (NML) formulation of the stochastic complexity contains two components: the maximized log likelihood and a component that may be interpreted as the parametric complexity of the model. The stochastic complexity for the data, relative to a suggested model, serves as a criterion for model selection (MS). The calculation of the stochastic complexity can be considered as an implementation of the Rissanen's minimum description length (MDL) principle. To obtain an NML based model selection criterion for the Gaussian linear regression, Rissanen constrains the data space properly. In article [1] we show that the NML criterion is not invariant with respect to the data constraints and we study the effect of the constraints on the selection criterion. We demonstrate that the Rissanen's methodology can be generalized, and we show that new forms of the NML criterion can be obtained by varying the shape of the ellipsoidal constraint. The resulting 'natural' extension, to the best of the authors knowledge, has not appeared in the literature previously. A special emphasis is placed on the performance of the criterion when collinearity is present in data.

In article [2] we provide a rigorous analysis for the criteria derived in [1], and we further extend the results in article [1] by discussing more general convex constraints and its special case, rhomboidal constraint. We also compare the new criteria against five state-of-the-art selection rules by conducting Monte Carlo simulations for families of models commonly used in statistics and signal processing. Additionally, for the eight criteria which

are tested, we report results on their predictive capabilities for real world data sets.

In [3] we consider a new model averaging (MA) estimation method called weighted average least squares (WALS) which was introduced by Magnus et al. in 2010. The WALS estimator has good risk profile and its computational burden is light. We demonstrate that the WALS technique works efficiently even when the number of regressors is huge whereas many other existing MA or MS techniques are infeasible. We provide the basic theory behind WALS and study its estimation capabilities with respect to backwards elimination technique when comparing the hip fracture treatments costs between hospital districts in Finland.

Article [4] continues our work which starts from the WALS framework but then we view estimation from the perspective of the the penalized least squares (PenLS) technique. We characterize a wide class of shrinkage estimators where WALS estimators belong, and we derive the efficiency bound for the shrinkage estimators. We demonstrate that shrinkage estimators can be obtained by using the PenLS technique. Then we derive sufficient conditions for the PenLS estimator to belong to the class of shrinkage estimators. The PenLS technique provides a convenient tool to implement MA estimators into practice. We show that many well-known estimators can be characterized by using the defining properties of shrinkage estimators. We compare the performance of various PenLS estimators with the performance of our benchmark, the Laplace estimator, using simulation experiments within a realistic set-up.

In article [5] we study estimation of a classical nonparametric regression model by employing an approximating Gaussian linear regression model. The main theme of article [5] is the problem of selecting the weights for averaging across estimates obtained from a set of models in Gaussian linear regression. Some existing MA methods are based on exponential AIC or BIC weights, and Bayesian MA is a related technique. In this article we introduce a new MA technique by selecting the model weights using Rissanen's NML criterion. We compare the performance of the alternative MA estimators in simulation experiments.

For spline smoothing we rewrite in [6] the smooth estimation as a linear mixed model (LMM) where the smoothing parameter appears as the ratio between the variance of the error terms and the variance of random effects. Smoothing methods that use basis functions with penalization can utilize the maximum likelihood (ML) theory in the LMM framework. We introduce the NML model selection criterion for LMMs and propose an automatic databased spline smoothing method that utilize the MDL criterion. Simulation study shows that the performance of MDL in spline smoothing is close to that of the BIC criterion.

Article [7] has two purposes. First, we develop a risk adjustment model for hip fracture mortality using logistic regression and examine the impact of the length of the register follow-up period on adjusting the performance indicator for three comorbidities: congestive heart failure, cancer and di-

abetes. All three comorbidities have an effect on hip fracture mortality. The results indicate that for congestive heart failure all available medical history should be used, while for cancer it is enough to use only records from half a year before the fracture. For diabetes the choice of time period is not as clear, but using records from three years before the fracture seems to be a reasonable choice. The second purpose is to introduce an implementation of the MDL principle for model selection in logistic regression. This is carried out by using the NML technique. However, the computational burden becomes too heavy to apply the usual NML criterion. The idea of sequential NML (sNML) is introduced in order to enable evaluating the criterion efficiently also for models with large number of covariates. The results obtained by using sNML are compared to the corresponding results given by the traditional AIC and BIC model selection criteria.

5.2 Author's contribution to the articles

The seven articles contained in this thesis are joint research between myself and different co-authors. Prof. Liski (University of Tampere) is co-author in all articles except in [2] and [7]. In all articles, except in [2], the author has prepared and implemented the program codes used in numerical examples and simulations, and carried out the computations. He has also produced all the figures and tables in the articles, except in [2]. Prof. Liski proposed the theme of the article [1] while the process of writing the article was a joint effort.

Article [2] is a continuation of [1]. The topic was proposed by me and Prof. Liski. The work process of preparing article [2] started with several months weekly meetings with Dr. Ciprian Doru Giurcaneanu (Tampere University of Technology) and Dr. Alireza Razavi (Tampere University of Technology) where we discussed the research problems relating to article [2], proposed various approaches to solving problems and went through tentative proofs of the results. My main contributions to [2] are in the propositions 3.1, 3.2, 3.3 and their proofs in section 3.

Article [3] is joint work with Prof. Liski, Dr. Reijo Sund (National Institute for Health and Welfare) and Ms. Merja Juntunen (National Institute for Health and Welfare). Ms. Juntunen provided preprocessing of the data and Dr. Sund gave his expertise on hip fracture data and their analysis. Prof. Liski brought the idea to study WALS estimation but otherwise [3] is the outcome of an interactive collaboration between me and Prof. Liski.

In article [4] we go on with the WALS estimation method but we widened our perspective to study estimation in the context of the penalized least squares (PenLS) and shrinkage estimation. The article is a joint work between me, Prof. Liski and Prof. Unto Häkkinen (National Institute for Health and Welfare). Prof. Häkkinen provided the data which is used in the simulation experiments. Prof. Liski proposed the PenLS approach, otherwise the article is a result of collaborative writing and discussion process, except sections 5 and 6 for which I was mainly responsible.

Articles [5] and [6] are joint works with Prof. Liski. In [5] Prof. Liski proposed using the NML criterion in model averaging and in [6] he proposed using NML in the linear mixed model framework. Here again the articles grew up as a result of a collaborative process similar to that of article [4]. However, I was mainly responsible for section 5 in [5] and section 1.5 in [6].

Article [7] is joint work between me, Prof. Ioan Tabus (Tampere University of Technology), Dr. Sund and Prof. Häkkinen. Prof. Tabus proposed the topic of sNML for the article and section 4 was prepared in collaboration with Prof. Tabus. Prof. Tabus also provided me guidance and help as my supervisor and helped in the preparation of the final version of the paper. Dr. Sund collaborated in the planning of analyses and interpretation of results. Prof. Häkkinen provided data for the analysis.

6

Conclusions

The research presented in this thesis is focused on three mutually related main points. First, we study the MDL principle in MS and especially its applications within the NML framework. Second, we consider MA methods in linear regression as an alternative to MS and focus on techniques which are also computationally efficient. Third, we demonstrate the use of the studied methods in practice and apply them to large hip fracture data sets.

The MDL model selection is a central theme in this thesis. In [1] we derive a new family of MDL model selection criteria in Gaussian linear regression by extending the Rissanen methodology for computing the parametric complexity. In [2] we further extend the idea introduced in [1] and provide a rigorous analysis. We also compare the derived criteria against certain established MS rules by conducting Monte Carlo simulations. MS criteria typically seek for balance between good fit and complexity but this extended methodology may serve as a technique to tailor MS criteria that detect also other features of the model such as multicollinearity, for example. However, the effects and interpretations of data constraints call for further enlightenment which is an interesting topic for further research.

Although computation of the parametric complexity in logistic regression is straightforward, in principle, the computational load becomes overwhelming in practice when the number of covariates is large. That was the problem we encountered in [7] when we modeled hip fracture mortality using logistic regression. Therefore we introduced a sequential NML model selection method which is computationally feasible. We derive in [6] the MDL model selection criterion for linear mixed models by extending the Rissanen renormalization technique for linear regression. The connection between linear mixed models and smoothing splines makes it possible to apply the MDL criterion for modeling with smoothing splines.

In a regression problem with many regressors, a popular method to reduce the dimensionality of the model is to carry out tests about regression coefficients sequentially. Data analysts often run procedures such as forward addition of variables, or backwards deletion of variables. In [3] and [4] we

start considering the MS problem as a testing problem but we convert it into a problem of weighted average LS estimation. Because we want to consider all possible subsets of auxiliary regressors, the number of alternative model candidates is huge when the number of regressors is large. Hence estimation of all model weights is computationally too heavy a task. Therefore, in [4] we develop a substitute for weight estimation which utilizes shrinkage estimation and penalized least squares. This methodology can be extended to generalized linear models, and currently we are working on this problem. The extension makes it possible to apply the method on the hip fracture mortality modeling considered in [7]. In [5] we use the MDL criterion to select model weights in function estimation which can be considered as an MA problem. The technique presented in [5] is not computationally feasible for the weighted average LS model averaging in regression.

Of course, our interest in the MDL principle and our wish to learn more about it acted as a stimulus to choose the MDL approach as the main theme. We soon realized that this approach provides tough conceptual and technical challenges when applying it to data. In fact, the starting point for the study problems has been rather practical. The papers [1], [2], [6] and [7] are closely related to the question: How do you compute the parametric complexity of your model? If you are not able to compute the parametric complexity, you cannot use the NML approach. In our first study (Liski 2005) on hip fracture treatment costs in Finland we used propensity score analysis. This analysis has certain limitations when comparing numerous costs simultaneously. This aroused our interest in MA in connection of this problem. These attempts again brought out computational problems, and a corresponding need to develop methodology and software.

In this thesis, we have not made any attempt to compare various MS or MA methods, in general. We have compared the MDL criterion with several established MS criteria such as AIC, BIC, Mallows' C_p and GCV by carrying out simulation experiments in certain specified settings. There is no such unifying message from these experiments that some method would be uniformly the best. In our seven papers there is only one sentence containing a value judgement in favor of the MDL principle. In article [7] we write "The NML distributions offer a philosophically superior approach for the model selection problem". It may be interpreted as an expression of enthusiasm for the MDL principle at the moment of writing but not as a scientific hypothesis.

Finally, we conclude this section by presenting the main outcomes of this work by topic.

The MDL principle: In Gaussian regression the parametric complexity in the NML criterion is not finite and therefore the data space has to be constrained appropriately. The constraint introduced by Rissanen yields the form of the NML criterion which is most widely known. In [1] we show that the choice of constraint has an effect on the criterion and we impose an ellipsoidal constraint on the data space and we study more closely three special cases of ellipsoidal constraints which all lead to com-

pletely new forms of the NML criterion. This effect of the constraint on the criterion has not been discussed in previous literature on NML. In [2] we provide a more rigorous and general analysis of the problem. We present a general family of convex constraints and study thoroughly the special case ellipsoidal constraints, introduced in [1], and a new rhomboidal constraint.

The NML criterion and spline smoothing: In [6] we derive the NML criterion for linear mixed models. By utilizing the connection between linear mixed models and smoothing splines, the NML criterion is used in spline smoothing. We present an automated data based spline smoothing method using this newly derived criterion. Based on simulation experiments, the performance of the NML criterion seems to be close to that of the BIC criterion.

The sNML criterion and logistic regression: As the amount of data and the number of covariates in logistic regression increase, the traditional NML criterion becomes computationally infeasible. In article [7] we present a new model selection criterion for logistic regression which is based on the sequentially normalized maximum likelihood (sNML). This criterion is shown to be applicable also in large datasets when there are plenty of covariates.

Model averaging: Most existing model averaging (MA) methods are based on estimation of all model weights using exponential Akaike information criterion (AIC) or Bayesian information criterion (BIC) weights, for example. In [5] we use the NML criterion in choosing of model weights. We provide a comparison of MDL with AIC, BIC and Mallow's C criteria using simulation experiments. The main message is that the performance of the MA estimators based on MDL, BIC and Mallow's criteria are pretty close to each others.

A common challenge for a regression analyst is the selection of the best subset from a set of predictor variables in terms of some specified criterion. If the number of predictors is m , say, then the number of competing models is 2^m , and consequently the computational burden to estimate all the model weights becomes soon too heavy when m is large. The idea in [4] is to convert estimation of 2^m model weights into estimation of m shrinkage factors with trivial computational burden. We define the class of shrinkage estimators in view of MA and show that these shrinkage estimators can be constructed and estimated using penalized least squares by putting appropriate restrictions on the penalty function. Utilizing the relationship between shrinkage estimation and parameter penalization, we are able to build up computationally efficient MA estimators which are easy to implement into practice. These estimators include some known contributions, like the non-negative garrote of Breiman, the lasso-type estimator of Tibshirani and the SCAD (smoothly clipped absolute deviation) estimator of Fan and Li. In the simulation experiments we assess the quality of estimators in terms of their RMSE. In this competition the winners were the SCAD and non-negative garrote but the Laplace estimator did almost as well.

Hip fracture treatment costs and hip fracture mortality: In article [3] we compared hip fracture treatment costs between seven largest hospital districts in Finland. We found statistically significant differences in treatment costs between the largest hospital districts. By cost the most significant auxiliary variables are age, waiting for operation over 2 days, Parkinson disease, alcohol abuse, hypertension and diabetes.

In article [7] we develop a risk adjustment model for hip fracture mortality. Our results indicate that for congestive heart failure we should use all medical history available to us, while for cancer it is enough to use only records from half a year before the fracture. For diabetes the message is not clear, but using records from three years before the fracture seems to be a reasonable choice. The results obtained by using a sliding window do not change our previous conclusions on the effect of different comorbidities. This suggests that there has not been any remarkable changes in covariate effects within the time period under consideration.

We were also able to distinguish how much of the change in codelength is due to the observations that become new indications of a comorbidity as we increase the time period that we look back in time. In congestive heart failure the fit of the whole data improves as we get new indications of that comorbidity. On the other hand, with cancer the model fits worse especially among the new cancer indications. Also this suggests that cancer's effect on mortality is quite different from that of congestive heart failure.

References

- Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In: Petrov, B. N. and Csaki, F. (eds.) *Second International Symposium on Information Theory*, Akademiai Kiado, Budapest, pp. 267–281.
- Barron, A. R., Rissanen, J. and Yu, B. (1998). The MDL principle in modeling and coding, *Special Issue of Information Theory to Commemorate 50 Years of Information Theory*, 44(6), pp. 2743–2760.
- Breiman, L. (1995). Better subset regression using nonnegative garrote, *Technometrics*, 37(4), pp. 373–384.
- Brumback, B. A., Ruppert, D. and Wand, M. B. (1999). Comment on Shively, Kohn and Wood, *Journal of the American Statistical Association*, 94(447), pp. 794–797.
- Buckland, S. T., Burnham, K. P. and Augustin, N. H. (1997). Model selection: An integral part of inference, *Biometrics*, 53(2), pp. 603–618.
- Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multi-model inference*, New York, Springer-Verlag.
- Claeskens, G. and Hjort, N. I. (2008). *Model selection and model averaging*, Cambridge, Cambridge University Press.
- Cover, T. and Thomas, J. (1991). *Elements of Information Theory*, New York: Wiley, Interscience.
- Cox, D. R. and McCullagh, P. (1982). Some aspects of analysis of covariance, *Biometrics*, 38(3), pp. 541–61.
- Danilov, D. and Magnus, J. R. (2004). On the harm that ignoring pretesting can cause, *Journal of Econometrics*, 122(1), pp. 27–46.
- Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage, *Biometrika*, 81(3), pp. 425–456.

- Einmahl, J. H. J., Kumar, K. and Magnus, J. R. (2011). Bayesian model averaging and the choice of prior, *CentER Discussion Paper*, No. 2011–003.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, 96(456), pp. 1348–1360.
- Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools, *Technometrics*, 35(2), pp. 109–148.
- Gissler, M. (1999). Routinely collected registers in Finnish health research. In: Alho, J. (Ed.) *Statistics, Registries, and Science Experiences from Finland*. Keuruu, Statistics Finland.
- Gissler, M. and Haukka, J. (2004). Finnish health and social welfare registers in epidemiological research, *Norsk Epidemiologi*, 14(1), pp. 113–120.
- Graven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation, *Numerische Mathematik*, 31, 377–403.
- Gruber, M. H. J. (1998). *Improving Efficiency by Shrinkage*, New York, Marcel Dekker Inc.
- Grünwald, P. (2007). *The minimum description length principle*, Cambridge, Massachusetts, MIT Press.
- Hannan, E. L., Magaziner, J., Wang, J. J., Eastwood, E. A., Silberzweig, S. B., Gilbert, M., Morrison, R. S., McLaughlin, M. A., Orosz, G. M. and Siu, A. L. (2001). Mortality and locomotion 6 months after hospitalization for hip fracture: risk factors and risk-adjusted hospital outcomes, *Journal of the American Medical Association* 285(21), pp. 2736–2742.
- Hansen, B. E. (2007). Least squares model averaging, *Econometrica* 75(4), pp. 1175–1189.
- Hansen, M. and Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association* 96(454), 746–774.
- Hastie, T., Tibshirani, R. and Friedman, J. (2001), *The Elements of Statistical Learning*, New York, Springer.
- Heithoff, K. A. and Lohr, K. N. (eds.) (1990). *Hip fracture: Setting priorities for effectiveness research*, Washington, DC: National Academy Press.
- Hietala, M. (2009). *On choosing the variables for the risk-adjustment model when comparing hip fracture treatment practices* (in Finnish). Masters thesis, Department of Mathematics and Statistics, University of Tampere.
- Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, 12(1), pp. 55–67.

- Horn, R. A. and Johnson, C. R. (1985). *Matrix analysis*, Cambridge, Cambridge University Press.
- Hosmer, D. W., and Lemeshow, S. (2000). *Applied Logistic Regression*, New York. Wiley.
- Iezzoni, L. I. (2003). Range of risk factors. In: Iezzoni LI (Ed.) *Risk adjustment for measuring health care outcomes*, 3rd edition, Health Administration Press, Chicago.
- James, W. and Stein, C. (1961). Estimation with quadratic loss, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, University of California Press, pp. 361–379.
- Judge, G. G. and Bock, M. E. (1978). *The statistical implications of pre-test and Stein-rule estimators in econometrics*, Amsterdam, North-Holland.
- Judge, G. G., Griffiths, W. E., Hill, R. C., Lutkepohl. H, and Lee, T. C. (1985). *The theory and practice of econometrics*, New York, Wiley.
- Kolmogorov, A. (1965). Three approaches to the quantitative definition of information. *Problems Inform. Transmission*, 1(1), 1–7.
- Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency, *Annals of Mathematical Statistics*, 22(1), 79-86.
- Leamer, E. E. (1978). *Specification Searches: Ad Hoc Inference with Non-experimental Data*, New York, Wiley.
- Li, M. and Vitányi, P. (1997). *An introduction to Kolmogorov complexity and its applications* (revised and expanded second ed.), New York, Springer-Verlag.
- Liang, F. and Barron, A. R. (2005). Exact minimax predictive density estimation and MDL. In: Grünwald, P.D., Myung, I.J. and Pitt, M.A. (Eds.). *Advances in Minimum Description Length: Theory and Applications*, Cambridge, MA: MIT Press.
- Liporace, F. A., Egol, K. A., Tejwani, N., Zuckerman, J. D. and Koval, K. J. (2005). Whats new in hip fractures? Current concepts, *The American Journal of Orthopedics*, 34(2), pp. 66-74.
- Magnus, J. R. (1999). The traditional pretest estimator, *Theory of Probability and Its Applications*, 44, pp. 293-308.
- Magnus, J. R. (2002). Estimation of the mean of a univariate normal distribution with a known variance, *Econometrics Journal*, 5(1), pp. 225-236.
- Magnus, J. R. and Durbin, J. (1999). Estimation of regression coefficients of interest when other regression coefficients are of no interest, *Econometrica*, 67(3), pp. 639–643.
- Magnus, J. R., Powell, O. and Prüfer, P. (2010). A comparison of two model averaging techniques with an application to growth empirics, *Journal of Econometrics*, 154(2), pp. 139–153.
- Mallows, C. L. (1973). Some comments on C_p , *Technometrics*, 15(4), pp. 661–675.

- McCulloch, C.E. and Searle, S.R. (2001). *Generalized, linear and mixed models*, New York, Wiley-Interscience.
- Miller, A. (2002). *Subset selection in regression*, Boca Raton, Chapman & Hall.
- Morris, C., Radhakrishnan, R. and Sclove, S.L. (1972). Nonoptimality of preliminary test estimators for the mean of a multivariate normals distribution. *Annals of Mathematical Statistics*, 43(5), pp. 1481–1490.
- Nelder, J. A. and Wedderburn, R. W. (1972). "Generalized linear models", *Journal of the Royal Statistical Society, Series A*, 135(3), pp. 370–384.
- Nurmi, I., Narinen, A., Lüthje, P. and Tanninen, S. (2003). Cost analysis of hip fracture treatment among the elderly for the public health services: a 1-year prospective study in 106 consecutive patients, *Archives of Orthopaedic and Trauma Surgery*, 123(10), pp. 551–554.
- Peltola, M., Juntunen, M., Häkkinen, U., Rosenqvist, G., Seppälä, T. and Sund, R. (2011). A methodological approach for register-based evaluation of cost and outcomes in health care, *Annals of Medicine*, 43(Suppl 1), pp. 4–13.
- Rao, C.R., Toutenburg, H., Shalabh and Heumann, C. (2008). *Linear Models and Generalizations, Least Squares and Alternatives*, 3rd ed., Springer-Verlag.
- Rissanen, J. (1978). Modeling by shortest data description, *Automatica*, 14(1), pp. 465–471.
- Rissanen, J. (1986). Stochastic complexity and modeling. *The Annals of Statistics* 14, 1080–1100.
- Rissanen, J. (1987). Stochastic complexity. *Journal of the Royal Statistical Society*, series B 49, 223–239. Discussion: pages 252–265.
- Rissanen, J. (1996). Fisher information and stochastic complexity, *Information Theory, IEEE Transactions on Information Theory*, 42(1), pp. 40–47.
- Rissanen, J. (2000). MDL denoising, *Information Theory, IEEE Transactions on*, 46(1), pp. 2537–2543.
- Rissanen, J. (2007). *Information and complexity in statistical modeling*, Springer, New York.
- Rissanen, J. (2012). *Optimal parameter estimation*, Cambridge, Cambridge University Press.
- Rissanen, J. and Roos, T. (2007). Conditional NML universal models, *In Proceedings 2007 Information Theory and Applications Workshop*, IEEE press, pp. 337–341.
- Robinson, C.M., Court-Brown, C.M., McQueen, M.M. and Christie, J. (1995). Hip fractures in adults younger than 50 years of age. Epidemiology and results. *Clinical Orthopaedics and Related Research*, 312, pp. 238–246.

- Roos, T. and Rissanen, J. (2008). On sequentially normalized maximum likelihood models. *Workshop on Information Theoretic Methods in Science and Engineering* (WITMSE-08), Tampere, Finland, August 18-20.
- Ruppert, D., Wand, M.P. and Carroll, R.J. (2003). *Semiparametric regression*, Cambridge, Cambridge University Press.
- Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics*, 6(2), pp. 461-464.
- Searle, S.R., Casella, G., and McCulloch, C.E. (1992). *Variance components*, New York, Wiley.
- Seber, G.A.F. (1977). *Linear regression analysis*, New York, Wiley.
- Seber, G.A.F. and Lee, A.J. (2003). *Linear regression analysis* (2nd edition), Wiley Interscience.
- SIGN (2002). Prevention and management of hip fracture in older people. Edinburgh: Scottish Intercollegiate Guidelines Network (SIGN).
- Shtarkov, Y.M. (1987). Universal sequential coding of single messages. (translated from Russian) *Problems of Information Transmission* 23(3), 3-17.
- Sund, R. (2006). Lonkkamurtumien ilmaantuvuus Suomessa 1998-2002, *Duodecim*, 122(9), pp. 1085-1091.
- Sund, R. (2008). *Methodological perspectives for register-based health system performance assessment. Developing a hip fracture monitoring system in Finland*, STAKES research report 174, Helsinki, National Research and Development Centre for Welfare and Health.
- Sund, R. (2012). Quality of the Finnish hospital discharge register: A systematic review, *Scandinavian Journal of Public Health*. 40(6), pp. 505-515.
- Sund, R., Juntunen, M., Lüthje, P., Huusko, T. and Häkkinen, U. (2011). Monitoring the performance of hip fracture treatment in Finland, *Annals of Medicine*, 43(Suppl 1), pp. 39-46.
- Sund, R., Nurmi-Lüthje, I., Lüthje, P., Tanninen, S., Narinen, A. and Keskimäki, I. (2007). Comparing properties of audit data and routinely collected register data in case of performance assessment of hip fracture treatment in Finland, *Methods of Information in Medicine*, 46(5), pp. 558-566.
- Sund, R., Juntunen, M., Lüthje, P., Huusko, T., Mäkelä, M., Linna, M., Liski, A. and Häkkinen, U. (2008). PERFECT - Lonkkamurtuma - Hoitoketjujen toimivuus, vaikuttavuus ja kustannukset lonkkamurtumapotilailla, STAKES, Työpapereita 18.
- Tabus, I. and Rissanen, J. (2006). Normalized Maximum Likelihood models for logit regression, In: Liski, E.P., Isotalo, J., Niemel, J., Puntanen, S., and Styan, G.P.H., (eds.) *Festschrift for Tarmo Pukkila on his 60th Birthday*, University of Tampere, Department of Mathematics, Statistics and Philosophy, Report A 368.

- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society B*, (58)1, pp. 267–288.
- Toro-Vizcarrondo, C. and Wallace, W. D. (1968). A test of the mean square error criterion for restrictions in linear regression, *Journal of the American Statistical Association*, 63, pp. 558–572.
- Zuckerman, J. D. (1996). Hip fracture, *The New England Journal of Medicine*, 334(23), pp. 1519–1525.

Original articles

[1] Liski, E. P. and Liski, A. (2009)

**Minimum description length model selection
in Gaussian regression under data constraints**

In: Schipp, B., Krämer, W. (eds.) *Statistical Inference, Econometric Analysis and Matrix Algebra, Festschrift in Honour of Götz Trenkler*, Springer, pp. 201–208.

Minimum description length model selection in Gaussian regression under data constraints

ERKKI P. LISKI and ANTTI LISKI
University of Tampere and
Tampere University of Technology, Finland

Abstract

The normalized maximum likelihood (*NML*) formulation of the stochastic complexity (Rissanen 1996) contains two components: the maximized log likelihood and a component that may be interpreted as the parametric complexity of the model. The stochastic complexity for the data, relative to a suggested model, serves as a criterion for model selection. The calculation of the stochastic complexity can be considered as an implementation of the minimum description length principle (*MDL*) (cf. Rissanen 2007). To obtain an *NML* based model selection criterion for the Gaussian linear regression, Rissanen (2000) constrains the data space appropriately. In this paper we demonstrate the effect of the data constraints on the selection criterion. In fact, we obtain various forms of the criterion by reformulating the shape of the data constraints. A special emphasis is placed on the performance of the criterion when collinearity is present in data.

2000 *Mathematics Subject Classification.* 62B10, 62J05, 62F99.

Key words or phrases. Stochastic complexity, Parametric complexity, Normalized maximum likelihood, Collinearity, Linear regression..

1 Introduction

The variable selection problem is most familiar in the Gaussian regression context. Suppose that the response variable \mathbf{y} and the potential explanatory variables $\mathbf{x}_1, \dots, \mathbf{x}_K$ are vectors of n observations. The problem of variable selection arises when one wants to decide which variables to include into the model. If we let γ index the subsets of $\mathbf{x}_1, \dots, \mathbf{x}_K$ and let k_γ be the size of the γ th subset, then the problem is to select and fit a model of the form

$$\mathbf{y} = \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma + \boldsymbol{\epsilon}, \quad (1)$$

where \mathbf{X}_γ is an $n \times k_\gamma$ regression matrix corresponding to the γ th subset, $\boldsymbol{\beta}_\gamma$ is the $k_\gamma \times 1$ vector of unknown regression coefficients and $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma_\gamma^2 \mathbf{I})$.

Let $\hat{\boldsymbol{\theta}}_\gamma = (\hat{\boldsymbol{\beta}}_\gamma, \hat{\sigma}_\gamma^2)$ denote the *ML* estimates

$$\hat{\boldsymbol{\beta}}_\gamma = (\mathbf{X}_\gamma' \mathbf{X}_\gamma)^{-1} \mathbf{X}_\gamma' \mathbf{y} \quad \text{and} \quad \hat{\sigma}_\gamma^2 = RSS_\gamma / n \quad (2)$$

of $\boldsymbol{\beta}_\gamma$ and σ_γ^2 from the model (1), where $RSS_\gamma = \|\mathbf{y} - \hat{\mathbf{y}}_\gamma\|^2$ is the residual sum of squares and $\hat{\mathbf{y}}_\gamma = \mathbf{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma$ is the vector of fitted values. Here we assume that \mathbf{X}_γ is of full column rank.

The two most well-known methods for model selection are the *Akaike information criterion* or *AIC* (Akaike 1973, Burnham and Anderson 2002) and the *Bayesian information criterion* or *BIC* (Schwarz 1978). The *Akaike information criterion* is defined by

$$AIC(\gamma) = -2 \log f(\mathbf{y}; \hat{\boldsymbol{\theta}}_\gamma) + 2k_\gamma,$$

where $f(\mathbf{y}; \boldsymbol{\theta}_\gamma)$ is the density function of \mathbf{y} . The corresponding *BIC* criterion is

$$BIC(\gamma) = -2 \log f(\mathbf{y}; \hat{\boldsymbol{\theta}}_\gamma) + k_\gamma \log n.$$

The *MDL* principle for statistical model selection is based on the idea to capture regular features in data by constructing a model in a certain class which permits the shortest description of the data and the model itself. Rissanen's (1996, 2007) *MDL* approach to modeling utilizes ideas of coding theory. The expression

$$-\log \hat{f}(\mathbf{y}; \gamma) = -\log f(\mathbf{y}; \hat{\boldsymbol{\theta}}_\gamma) + \log C(\gamma) \quad (3)$$

defines the "shortest code length" for the data \mathbf{y} that can be obtained with the model γ and it is called the *stochastic complexity* of \mathbf{y} , given γ .

Under certain conditions $\log C(\gamma)$ has the estimate (Rissanen 1996)

$$\log C(\gamma) = \log \frac{n}{2\pi} + \log \int |\mathbf{J}(\boldsymbol{\theta}_\gamma)|^{1/2} d\boldsymbol{\theta}_\gamma + o(1), \quad (4)$$

where $|\mathbf{J}(\boldsymbol{\theta}_\gamma)|$ is the determinant of the Fisher's information matrix. Since the last term $o(1)$ in (4) goes to zero as $n \rightarrow \infty$ and the second term is constant, asymptotically $\log C(\gamma)$ behaves like the first term. Thus we see the asymptotic connection with the *BIC*. For some important models $\log C(\gamma)$ can be calculated exactly, for example by using the *NML* technique. In statistical literature the *MDL* principle is often confused with a particular implementation of it as the selection criterion *BIC* (For discussion see Grünwald 2007 p. 552). In fact, the stochastic complexity (3) has the adaptation property that it behaves more like *AIC* when the number of parameters is getting large compared with the number of observations.

2 Selection by stochastic complexity

Assume that \mathbf{y} follows the Gaussian linear model (1) with $\boldsymbol{\theta}_\gamma = (\boldsymbol{\beta}_\gamma, \sigma_\gamma^2)$. Here we consider the family of models

$$\mathcal{M}_\gamma = \{f(\mathbf{y}; \boldsymbol{\theta}_\gamma) : \gamma \in \Gamma\} \quad (5)$$

defined by the normal densities $f(\mathbf{y}; \boldsymbol{\theta}_\gamma)$, where Γ denotes a set of subsets of $\mathbf{x}_1, \dots, \mathbf{x}_K$, i.e. the set of models we wish to consider.

After observing \mathbf{y} we may determine the maximum likelihood (*ML*) estimate $\hat{\boldsymbol{\theta}}_\gamma = \hat{\boldsymbol{\theta}}_\gamma(\mathbf{y})$ of $\boldsymbol{\theta}_\gamma$ such that $f(\mathbf{y}; \hat{\boldsymbol{\theta}}_\gamma) = \max_{\boldsymbol{\theta}_\gamma} f(\mathbf{y}; \boldsymbol{\theta}_\gamma)$. Rissanen (1996) introduced the *NML* function

$$\hat{f}(\mathbf{y}; \gamma) = \frac{f(\mathbf{y}; \hat{\boldsymbol{\theta}}_\gamma(\mathbf{y}))}{C(\gamma)} \quad \text{with} \quad C(\gamma) = \int f(\mathbf{y}; \hat{\boldsymbol{\theta}}_\gamma(\mathbf{y})) d\mathbf{y}, \quad (6)$$

where $\hat{f}(\mathbf{y}; \gamma)$ is a density function, provided that $C(\gamma)$ is bounded. The *NML* density function provides a general technique to apply the *MDL* (minimum description length) principle. Therefore the derivation of the *NML* density is a crucial step in the practical implementation of the *MDL* principle.

For each model $\gamma \in \Gamma$ we have an *NML* density (6) which depends on γ . In the sequel, $\hat{f}(\mathbf{y}; \gamma)$ refers to the *NML* density of the model γ , and $C(\gamma)$ denotes the corresponding normalizing constant. Now the stochastic complexity (3) can be calculated by using the *NML* density:

$$-\log \hat{f}(\mathbf{y}; \gamma) = -\log f(\mathbf{y}; \hat{\boldsymbol{\theta}}_\gamma(\mathbf{y})) + \log C(\gamma).$$

The last term in the equation (3) is called *the parametric complexity* of the model. According to the *MDL* principle we seek to find the index value $\gamma = \hat{\gamma}$ that minimizes the stochastic complexity (3). The basics of the *MDL* theory are presented in the recent books by Grünwald (2007) and by Rissanen (2007).

Since the following development will be for a fixed γ , we may drop the subindex γ for a while without loss of clarity. It turns out that the *NML* function (6) for the normal distribution is undefined, since the normalizing constant C is not bounded. Hence Rissanen (2000) suggested the constrained data space

$$\mathcal{Y}(s, R) = \{\mathbf{y} : \hat{\boldsymbol{\beta}}' \mathbf{X}' \mathbf{X} \hat{\boldsymbol{\beta}} \leq nR, \hat{\sigma}^2 \geq s\}, \quad (7)$$

where $s > 0$ and $R > 0$ are given positive constants. Then the *NML* density under the constraints (7) will be

$$\hat{f}(\mathbf{y}; s, R) = f(\mathbf{y}; \hat{\boldsymbol{\theta}}(\mathbf{y})) / C(s, R), \quad (8)$$

where now the normalizing constant $C(s, R)$ depends on two hyperparameters s and R .

To get rid of these hyperparameters Rissanen (2000) applied another level of normalization. Maximizing the function (8) with respect of R and s yields the *ML* estimates $\hat{R} = \|\hat{\mathbf{y}}\|^2 / n$ and $\hat{s} = \hat{\sigma}^2$. The maximized *NML* function *mNML* is obtained by substituting these estimates into (8) in place of s and R . Then the function *mNML* is normalized. In this second stage normalization the data space is constrained such that

$$\mathcal{Y} = \{\mathbf{y} : nR_1 \leq \|\hat{\mathbf{y}}\|^2 \leq nR_2, s_1 \leq \hat{\sigma}^2 \leq s_2\}, \quad (9)$$

where $0 < R_1 < R_2$ and $0 < s_1 < s_2$ are given positive constants. By normalizing the function $\hat{f}(\mathbf{y}; \hat{s}, \hat{R})$ we obtain the normalized $mNML$ function $\hat{f}(\mathbf{y})$, say. Finally the stochastic complexity (3) takes the form

$$-\log \hat{f}(\mathbf{y}) = \frac{n-k}{2} \log \hat{\sigma}^2 + \frac{k}{2} \log \hat{R} - \log \Gamma\left(\frac{n-k}{2}\right) - \log \Gamma\left(\frac{k}{2}\right) + c, \quad (10)$$

where $\Gamma(\cdot)$ denotes the gamma function and $c = \frac{n}{2} \log(n\pi) + \log[\log \frac{s_2}{s_1} \log \frac{R_2}{R_1}]$ is the same for all models, and hence it can be ignored. More details can be found in Rissanen (2000, 2007).

3 The effect of data constraints

For the Gaussian density $f(\mathbf{y}; \boldsymbol{\theta})$ the numerator in (8) takes a simple form

$$f(\mathbf{y}; \hat{\boldsymbol{\theta}}) = (2\pi\hat{\sigma}^2 e)^{-\frac{n}{2}},$$

but the normalizing constant $C(s, R)$ will essentially depend on two hyperparameters s and R . The estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ is a sufficient statistic for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ under the model (1). By sufficiency the density $f(\mathbf{y}; \boldsymbol{\theta})$ belonging to the family (5) can be written as

$$f(\mathbf{y}; \boldsymbol{\theta}) = f(\mathbf{y}|\hat{\boldsymbol{\theta}})g(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}), \quad (11)$$

where the conditional density $f(\mathbf{y}|\hat{\boldsymbol{\theta}})$ does not depend on the unknown parameter vector $\boldsymbol{\theta}$. The ML estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$, given in (2), are independent. Therefore

$$g(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2; \boldsymbol{\beta}, \sigma^2) = g_1(\hat{\boldsymbol{\beta}}; \boldsymbol{\beta}, \sigma^2)g_2(\hat{\sigma}^2; \sigma^2), \quad (12)$$

where $g_1(\hat{\boldsymbol{\beta}}; \boldsymbol{\beta}, \sigma^2)$ and $g_2(\hat{\sigma}^2; \sigma^2)$ are the densities of the ML estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$, respectively. Substituting $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ into (12) in place of $\boldsymbol{\beta}$ and σ^2 , respectively, yields (cf. Rissanen 2000 and 2007, p. 115)

$$g_1(\hat{\boldsymbol{\beta}}; \hat{\boldsymbol{\beta}}, \hat{\sigma}^2)g_2(\hat{\sigma}^2; \hat{\sigma}^2) = A_{n,k}(\hat{\sigma}^2)^{-\frac{k}{2}-1}, \quad (13)$$

where

$$A_{n,k} = \frac{|\mathbf{X}'\mathbf{X}|^{1/2} \left(\frac{n}{2e}\right)^{n/2}}{(2\pi)^{k/2} \Gamma\left(\frac{n-k}{2}\right)}.$$

Utilizing the factorization (11) and the result (13) we get the normalizing constant $C(s, R)$ under the constraint (7) corresponding to (8) as follows:

$$\begin{aligned} C(s, R) &= \int_{\mathcal{T}(s,R)} \left[\int_{\mathcal{Y}(\hat{\boldsymbol{\theta}})} f(\mathbf{y}|\hat{\boldsymbol{\theta}}) d\mathbf{y} \right] \bar{g}(\hat{\sigma}^2) d\hat{\boldsymbol{\theta}} \\ &= A_{n,k} \int_s^\infty (\hat{\sigma}^2)^{-\frac{k}{2}-1} d\hat{\sigma}^2 \int_{\mathcal{B}(R)} d\hat{\boldsymbol{\beta}} \end{aligned}$$

$$= A_{\nu,k} V_k \frac{2}{k} \left(\frac{R}{s} \right)^{k/2}, \quad (14)$$

where $\mathcal{T}(s, R) = \{\hat{\boldsymbol{\theta}} : \hat{\sigma}^2 \geq s, \hat{\boldsymbol{\beta}}' \mathbf{Q} \hat{\boldsymbol{\beta}} \leq nR\}$ and \mathbf{Q} is a $k \times k$ positive definite matrix. Integrating the inner integral in the first line of (14) over $\mathcal{Y}(\hat{\boldsymbol{\theta}}) = \{\mathbf{y} : \hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{y})\}$ for a fixed value of $\hat{\boldsymbol{\theta}}$ gives unity. In the last line of (14)

$$V_k R^{k/2} = \frac{\pi^{k/2} n R^{k/2}}{\frac{k}{2} \Gamma(\frac{k}{2}) |\mathbf{Q}|^{1/2}}$$

is the volume of an ellipsoid

$$\mathcal{B}(\mathbf{Q}, R) = \{\hat{\boldsymbol{\beta}} : \hat{\boldsymbol{\beta}}' \mathbf{Q} \hat{\boldsymbol{\beta}} \leq nR\} \quad (15)$$

(cf. Cramer, p. 120).

The form of the stochastic complexity under the ellipsoidal constraint (15) takes the form

$$-\log \hat{f}(\mathbf{y}) = \frac{n-k}{2} \log \hat{\sigma}^2 + \frac{k}{2} \log \hat{R} - \log \Gamma\left(\frac{n-k}{2}\right) - \log \Gamma\left(\frac{k}{2}\right) + \frac{1}{2} \log \frac{|\mathbf{X}'\mathbf{X}|}{|\mathbf{Q}|}, \quad (16)$$

where $\hat{R} = \hat{\boldsymbol{\beta}}' \mathbf{Q} \hat{\boldsymbol{\beta}} / n$. The constant c , given in (10), is not essential in model comparison, and hence it is omitted. If we choose the constraint $\mathcal{B}(\mathbf{X}'\mathbf{X}, R)$ in (15), then $\log \frac{|\mathbf{X}'\mathbf{X}|}{|\mathbf{Q}|} = 0$ and the stochastic complexity (16) takes the form (10). This is the constraint Rissanen (2000 and 2007) uses. It is now clear that the matrix \mathbf{Q} in the ellipsoidal constraint (15) has an essential effect on the stochastic complexity.

4 Effects of collinearity

If we apply Stirling's approximation

$$\Gamma(x+1) \approx (2\pi)^{1/2} (x+1)^{x+1/2} e^{-x-1}$$

to Γ -functions in (16), omit the terms that do not depend on γ or k_γ and multiply (16) by 2, just for convenience, we have the *NML* criterion function of the form

$$MDL(\gamma, \mathbf{Q}) = n \log S_\gamma^2 + k_\gamma \log F(\mathbf{Q})_\gamma + \log[k_\gamma(n - k_\gamma)] + \log \frac{|\mathbf{X}'_\gamma \mathbf{X}_\gamma|}{|\mathbf{Q}|}, \quad (17)$$

where

$$S_\gamma^2 = \frac{RSS_\gamma}{n - k_\gamma} \quad \text{and} \quad F(\mathbf{Q})_\gamma = \frac{\hat{\boldsymbol{\beta}}'_\gamma \mathbf{Q} \hat{\boldsymbol{\beta}}_\gamma}{k_\gamma S_\gamma^2}.$$

In the special case $\mathbf{Q} = \mathbf{X}'\mathbf{X}$ the criterion (17) takes the form

$$MDL(\gamma, \mathbf{X}'\mathbf{X}) = n \log S_\gamma^2 + k_\gamma \log F_\gamma + \log[k_\gamma(n - k_\gamma)], \quad (18)$$

where

$$F_\gamma = \frac{\mathbf{y}'\mathbf{y} - RSS_\gamma}{kS_\gamma^2}$$

is the usual F -statistic. The formulation (18) was presented in Liski (2006), and also Hansen and Yu (2001) considered it in the context of a slightly different criterion.

Consider the set of models \mathcal{M}_k , where $k = k_\gamma$ and $RSS = RSS_\gamma$ for all $\gamma \in \mathcal{M}_k$. Then clearly the criterion (18) does not discriminate the models in \mathcal{M}_k . Assume that we have a satisfactory set of explanatory variables $\{x_1, \dots, x_{k-1}\}$ and we try to add new variables x_k and x_{k+1} . Consider a situation when both the model $\{x_1, \dots, x_{k-1}, x_k\}$, say γ_1 , and $\{x_1, \dots, x_{k-1}, x_{k+1}\}$, say γ_2 , yield the same, or a very close, residual sum of squares RSS , i.e. the models lie in \mathcal{M}_k . Hence, in terms of the MDL criterion (18), the two models are indistinguishable.

Assume that due to the collinearity between x_1, \dots, x_{k-1}, x_k , for example, the model yields large standard errors and low t -statistics for the estimates of the regression coefficients. On the other hand, the model with explanatory variables $x_1, \dots, x_{k-1}, x_{k+1}$ may still have satisfactory t -statistics. Clearly, this second model would be better, if our interest is also in regression coefficients, not only in prediction. However, the MDL criterion (18) fails to identify it. Note that AIC and BIC criteria have this same property.

For a collinear model γ the determinant $|\mathbf{X}'_\gamma \mathbf{X}_\gamma| \approx 0$ and the ML estimates of the regression coefficients become unstable, which may lead to a large value of $\|\hat{\beta}_\gamma\|^2$ (cf. Belsley 1991, for example). Let us further consider the set of models \mathcal{M}_k and take $\mathbf{Q} = \mathbf{I}$ in (17). Then in the criterion $MDL(\gamma, \mathbf{I})$

$$F(\mathbf{I})_\gamma = \frac{\|\hat{\beta}_\gamma\|^2}{kS_\gamma^2}$$

and the last term in (17) is $\log |\mathbf{X}'_\gamma \mathbf{X}_\gamma|$. Due to collinearity, $\log |\mathbf{X}'_{\gamma_1} \mathbf{X}_{\gamma_1}| < \log |\mathbf{X}'_{\gamma_2} \mathbf{X}_{\gamma_2}|$, but on the other hand $\|\hat{\beta}_{\gamma_1}\|^2$ tends to be larger than $\|\hat{\beta}_{\gamma_2}\|^2$. Thus the criterion (17) with $\mathbf{Q} = \mathbf{I}$ responds to the collinearity, but the message is not quite clear, since the two terms have opposite effects. If we use the criterion $MDL(\gamma, (\mathbf{X}'\mathbf{X})^2)$, then

$$F((\mathbf{X}'\mathbf{X})^2)_\gamma = \frac{\|\mathbf{X}'_\gamma \mathbf{y}\|^2}{kS_\gamma^2}$$

and the last term in (17) is $-\log |\mathbf{X}'_\gamma \mathbf{X}_\gamma|$. Now clearly the last term penalises the collinearity.

An example: STEAM data

As an example we consider the STEAM data set (Draper and Smith 1981, p. 616; Miller p. 69) which contains 25 observations on 10 variables. The response y is *pounds of steam used monthly* (the variable 1 in Draper and Smith), and the other 9 variables constitute the set of potential explanatory variables.

Table 1: Five best-fitting subsets of two and three variables, and two models of four variables for the STEAM data.

Variables	RSS_γ	$MDL(\gamma, \mathbf{X}'\mathbf{X})$	$MDL(\gamma, \mathbf{I})$	$MDL(\gamma, (\mathbf{X}'\mathbf{X})^2)$
1, 7	8.93	4.251	-0.818	10.618
5, 7	9.63	5.904	0.611	12.318
2, 7	9.78	6.258	1.226	12.631
4, 7	15.60	16.511	11.342	22.893
7, 9	15.99	17.051	11.680	23.486
4, 5, 7	7.34	7.744	-0.278	17.357
1, 5, 7	7.68	8.696	-0.066	18.977
1, 7, 9	8.61	11.087	2.847	21.221
1, 4, 7	8.69	11.283	3.276	21.011
5, 7, 8	8.71	11.321	3.121	21.291
2, 4, 5, 7	7.162	14.671	-18.112	-15.699
1, 2, 5, 7	7.156	14.656	-3.367	-0.954

We center and scale the explanatory variables which does not affect the fitted model but $\mathbf{X}'_\gamma \mathbf{X}_\gamma$ is the correlation matrix. Here the $MDL(\gamma, \mathbf{X}'\mathbf{X})$ increases monotonously as the function of RSS_γ when $k_\gamma = k$ is fixed. However, $MDL(\gamma, \mathbf{I})$ and $MDL(\gamma, (\mathbf{X}'\mathbf{X})^2)$ respond to collinearity. The two and three variable sets of explanatory variables given in Table 1 are not collinear. Therefore also $MDL(\gamma, \mathbf{I})$ and $MDL(\gamma, (\mathbf{X}'\mathbf{X})^2)$ put the models almost in same order as $MDL(\gamma, \mathbf{X}'\mathbf{X})$. However, the four variable models $\{x_2, x_4, x_5, x_7\}$ and $\{x_1, x_2, x_5, x_7\}$ have practically the same value of $MDL(\gamma, \mathbf{X}'\mathbf{X})$, but both $MDL(\gamma, \mathbf{I})$ and $MDL(\gamma, (\mathbf{X}'\mathbf{X})^2)$ strongly prefer $\{x_2, x_4, x_5, x_7\}$ to $\{x_1, x_2, x_5, x_7\}$. This is because the variables x_1, x_2, x_5, x_7 are much more collinear (the determinant of the correlation matrix $|\mathbf{R}| = 0.033$) than the variables x_2, x_4, x_5, x_7 ($|\mathbf{R}| = 0.299$). The estimate vector $\|\hat{\beta}\|$ has larger value for the model $\{x_1, x_2, x_5, x_7\}$ than for $\{x_2, x_4, x_5, x_7\}$ which has an effect on the criterion $MDL(\gamma, \mathbf{I})$. Especially the size of the coefficient $\hat{\beta}_2$ and the intercept increase dramatically whereas the coefficients $\hat{\beta}_5$ and $\hat{\beta}_7$ remain practically same.

References

- Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. Pages 267-281 in B. N. Petrov, and F. Csaki, (eds.) *Second International Symposium on Information Theory*. Akademiai Kiado, Budapest.
- Belsley, H. (1991). *Conditioning Diagnostics*. Wiley: New York.
- Burnham, K. P. and Anderson D. R. (2002). *Model Selection and Multi-model Inference*. Springer-Verlag: New York.
- Cramer, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press: Princeton.

- Draper, N.R., and Smith, H. (1981). *Applied regression analysis, 2nd ed.*. Wiley: New York.
- Grünwald, P. D. (2007). *The minimum description length principle*. MIT Press: London.
- Hansen, A. J. and Yu, B. (2001). Model Selection and the Principle of Minimum Description Length. *Journal of the American Statistical Association*, 96, 746-774.
- Liski, E. P. (2006). Normalized ML and the MDL Principle for Variable Selection in Linear Regression. In: *Festschrift for Tarmo Pukkila on His 60th Birthday*, 159-172, Tampere, Finland.
- Miller, A. (2002). *Subset Selection in Regression, 2nd edition*. Chapman & Hall/CRC: New York.
- Rissanen, J. (1996). Fisher Information and Stochastic Complexity. *IEEE Trans. on Information Theory*, IT-42, 1, 40-47.
- Rissanen, J. (2000). MDL Denoising. *IEEE Trans. on Information Theory*, IT-46, 1, 2537-2543.
- Rissanen, J. (2007). *Information and Complexity and in Statistical Modeling*. Springer-Verlag: New York.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.

[2] Giurcaneanu, C. D., Razavi, S. A. and Liski, A. (2011)
**Variable selection in linear regression: Several
approaches based on normalized maximum
likelihood,**
Signal Processing, 91(8), pp. 1671–1692.



Review

Variable selection in linear regression: Several approaches based on normalized maximum likelihood

Ciprian Doru Giurcăneanu*, Seyed Alireza Razavi, Antti Liski

Department of Signal Processing, Tampere University of Technology, P.O. Box 553, FIN-33101 Tampere, Finland

ARTICLE INFO

Article history:

Received 2 June 2010

Received in revised form

14 March 2011

Accepted 21 March 2011

Available online 1 April 2011

Keywords:

Gaussian linear regression

Model selection

Normalized maximum likelihood

Rhomboidal constraint

Ellipsoidal constraint

ABSTRACT

The use of the normalized maximum likelihood (NML) for model selection in Gaussian linear regression poses troubles because the normalization coefficient is not finite. The most elegant solution has been proposed by Rissanen and consists in applying a particular constraint for the data space. In this paper, we demonstrate that the methodology can be generalized, and we discuss two particular cases, namely the rhomboidal and the ellipsoidal constraints. The new findings are used to derive four NML-based criteria. For three of them which have been already introduced in the previous literature, we provide a rigorous analysis. We also compare them against five state-of-the-art selection rules by conducting Monte Carlo simulations for families of models commonly used in signal processing. Additionally, for the eight criteria which are tested, we report results on their predictive capabilities for real life data sets.

© 2011 Elsevier B.V. All rights reserved.

Contents

1. Introductory remarks and problem formulation	1672
2. Parametric complexity with constraints	1673
2.1. General case	1673
2.2. Rissanen formula	1674
2.3. Rhomboidal constraint	1674
3. Ellipsoidal constraint	1675
3.1. Formulas from [18]	1675
3.2. Penalty terms	1675
3.3. Comparison of the penalty terms when two nested models are tested	1676
4. Experimental results	1678
4.1. Model selection criteria used in experiments	1678
4.2. Numerical examples	1678
5. Conclusions	1687
Acknowledgments	1688
Appendix A. Evaluation of the normalized maximum likelihood	1688

* Corresponding author. Tel.: +358 3 3115 3832; fax: +358 3 3115 4989.

E-mail addresses: ciprian.giurcaneanu@tut.fi (C.D. Giurcăneanu), alireza.razavi@tut.fi (S.A. Razavi), antti.liski@tut.fi (A. Liski).

Appendix B. Proofs of the main results within Section 3.2	1689
Appendix C. Proofs of the main results within Section 3.3	1690
References	1692

1. Introductory remarks and problem formulation

One of the fundamental research topics addressed in signal processing is the linear least-squares regression problem. Let the measurements $\mathbf{y} \in \mathbb{R}^{n \times 1}$ be modeled by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (1)$$

where $\mathbf{X} \in \mathbb{R}^{n \times m}$ is the regressor matrix having more rows than columns ($n > m$), $\boldsymbol{\beta} \in \mathbb{R}^{m \times 1}$ is the vector of unknown parameters, and the entries of $\boldsymbol{\varepsilon} \in \mathbb{R}^{n \times 1}$ are samples from an independent and identically distributed (i.i.d.) Gaussian process of zero-mean and variance τ . Hereafter, we denote vectors by boldface lowercase letters and matrices by boldface uppercase letters. The identity matrix of appropriate dimension is denoted by \mathbf{I} , while $\mathbf{0}$ denotes a null vector/matrix of appropriate dimension.

Because in most of the practical applications, not all the parameters β_1, \dots, β_m are equally important in modeling \mathbf{y} , one wants to eliminate those that are deemed to be irrelevant. This reduces to choose a subset of the regressor variables indexed by $\gamma \subseteq \{1, \dots, m\}$. It is customary to select γ by using either the Akaike Information Criterion (AIC) [1], or the Bayesian Information Criterion (BIC) [29]. Both AIC and BIC can be seen like particular cases of a more general class of asymptotic criteria which are expressed as the sum of two terms: the first one is given by the minus maximum log-likelihood, and the second one is a penalty coefficient that depends on the number of parameters and, in some cases, on the sample size [36, Appendix C].

It is widely recognized that BIC is equivalent with an information theoretic criterion called MDL (minimum description length) [23]. However, MDL is not only a simple formula, but it is a principle [8].

To show how the most recent MDL-based developments can be applied to the linear regression problem, we focus on the computation of the stochastic complexity (SC) [25,26]. Let $\boldsymbol{\beta}_\gamma \in \mathbb{R}^{k \times 1}$ be the vector of the unknown regression coefficients within the γ -subset. We denote the cardinality of γ by k , and we make the assumption that k is strictly positive. The case $k=0$ will be treated separately. The matrix \mathbf{X}_γ is given by the columns of \mathbf{X} that correspond to the γ -subset. Similarly with (1), we have

$$\mathbf{y} = \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma + \boldsymbol{\varepsilon}_\gamma, \quad (2)$$

where the entries of $\boldsymbol{\varepsilon}_\gamma$ are Gaussian distributed with zero-mean and unknown variance τ_γ . Under the hypothesis that \mathbf{X}_γ has full-rank, the maximum likelihood (ML) estimates are [31]: $\hat{\boldsymbol{\beta}}_\gamma(\mathbf{y}) = (\mathbf{X}_\gamma^\top \mathbf{X}_\gamma)^{-1} \mathbf{X}_\gamma^\top \mathbf{y}$ and $\hat{\tau}_\gamma(\mathbf{y}) = \|\mathbf{y} - \mathbf{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma\|^2 / n$, where the superscripts $(\cdot)^\top$ and $(\cdot)^{-1}$ denote the transpose and the matrix inverse, respectively. The operator $\|\cdot\|$ is employed for the Euclidean norm. Whenever it is clear from the context which measurements are used for estimation, the simpler notation $\hat{\boldsymbol{\beta}}_\gamma$ will be preferred to $\hat{\boldsymbol{\beta}}_\gamma(\mathbf{y})$. The same applies for the use of $\hat{\tau}_\gamma$ instead of $\hat{\tau}_\gamma(\mathbf{y})$. To evaluate the SC for the data vector \mathbf{y} ,

given the γ -structure, we have to compute

$$\text{SC}(\mathbf{y}; \gamma) = L(\mathbf{y}; \gamma) + L(a, b) + \frac{n}{2} \ln(n\pi), \quad (3)$$

$$L(\mathbf{y}; \gamma) = \frac{n-k}{2} \ln \hat{\tau}_\gamma + \frac{k}{2} \ln \frac{\|\mathbf{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma\|^2}{n} - \ln \Gamma\left(\frac{n-k}{2}\right) - \ln \Gamma\left(\frac{k}{2}\right), \quad (4)$$

$$L(a, b) = 2 \ln \ln \frac{b}{a}, \quad (5)$$

where $\ln(\cdot)$ denotes the natural logarithm and $\Gamma(\cdot)$ is the Euler integral of the second kind. Additionally, the real-valued hyper-parameters a and b satisfy the condition: $b > a$.

The complete formula includes also the description length for the γ -structure, $L(\gamma)$, whose expression is given in [26]. Because in many practical problems, the term $L(\gamma)$ has a marginal effect, we will ignore it. Example 4 in Section 4 will be the only case when we will consider the contribution of this term. For clarifications on the role of $L(\gamma)$, see [27].

The case $k=0$ is equivalent to $\gamma = \emptyset$, and occurs when the observations \mathbf{y} are assumed to be pure Gaussian noise with zero-mean and unknown variance. In this situation, the stochastic complexity takes the particular form

$$\text{SC}(\mathbf{y}; \emptyset) = L(\mathbf{y}; \emptyset) + \frac{1}{2} L(a, b) + \frac{n}{2} \ln(n\pi), \quad (6)$$

$$L(\mathbf{y}; \emptyset) = \frac{n}{2} \ln \frac{\|\mathbf{y}\|^2}{n} - \ln \Gamma\left(\frac{n}{2}\right), \quad (7)$$

where $L(a, b)$ is defined in (5). In this work, we neglect the terms given by $L(a, b)$. We refer to [7,26, Section 9.3] for a more elaborated discussion on the conditions when $2 \ln \ln(b/a)$ and $\ln \ln(b/a)$ can be dropped from (3) and (6), respectively.

In line with the MDL principle, selection of the best structure amounts to evaluate $\text{SC}(\mathbf{y}; \gamma)$ for all $\gamma \subseteq \{1, \dots, m\}$, and then to pick-up the subset that minimizes the stochastic complexity. Another information theoretic criterion which is akin to formulas in (3)–(5) and (6)–(7) has been derived in [9,10] by using a universal mixture model. More interestingly, Kay has proposed in [16] a selection rule based on exponentially embedded families (EEF) of probability density functions, and which is similar to the one introduced by Rissanen in [25]. A comparison of the criteria from [10,16,25], for the case when the noise variance is assumed to be known, can be found in [6, Section 3.3]. The minimum message length (MML) principle was recently used in [28] to yield two new model selection criteria, and it turned out that both of them are closely related to SC.

The fact that formulas which are almost the same with the expression of SC can be obtained by various approaches is indeed an indicator for the practitioner that the use of SC might be the right choice. However, for the

work presented in this paper, the most important is not the SC formula, but the methodology applied by Rissanen for its derivation. The central role is played by the normalized maximum likelihood (NML) density function:

$$\hat{f}(\mathbf{y}; \gamma) = \frac{f(\mathbf{y}; \hat{\boldsymbol{\beta}}_\gamma(\mathbf{y}), \hat{\tau}_\gamma(\mathbf{y}))}{C(\gamma)}, \quad (8)$$

$$C(\gamma) = \int f(\mathbf{y}; \hat{\boldsymbol{\beta}}_\gamma(\mathbf{y}), \hat{\tau}_\gamma(\mathbf{y})) d\mathbf{y}, \quad (9)$$

where $f(\mathbf{y}; \hat{\boldsymbol{\beta}}_\gamma(\mathbf{y}), \hat{\tau}_\gamma(\mathbf{y}))$ is the ML. In the equation above, the domain of integration is the entire space of observations. Note also that (9) gives the definition of the *parametric complexity*. It was shown in [25,26] that NML has two important optimality properties which recommend it to be used in the evaluation of SC. More precisely, SC is computed as the code length associated with NML: $SC(\mathbf{y}; \gamma) = -\ln \hat{f}(\mathbf{y}; \gamma)$. The key point is that the parametric complexity in the linear regression case is not finite, or equivalently, the integral in (9) is not finite. To circumvent this difficulty, Rissanen proposed in [25] to constrain the integration domain in the space of observations such that the integral becomes finite, and this led to the criterion given by the formulas in (3)–(5) and (6)–(7).

We note in passing that, according to Scopus, Ref. [25] has been cited more than 50 times. Hence, the SC-criterion is widely used, and one of the reasons is the following. The criterion is independent of arbitrarily selected hyper-parameters if the terms that involve $L(a, b)$ are neglected. Surprisingly, for about one decade, it was totally ignored the important fact that the closed-form expression of the criterion depends on the particular constraint which has been involved in its derivation. Only recently, it was shown in [18] that two other criteria can be obtained by employing constraints which are different of the one used in [25].

The most recent findings lead to the conclusion that novel NML-based criteria can be devised by enforcing various constraints. However, in the previous literature, it was not investigated how the selection of the constraint influences the performance of the resulting criterion. To fill the gap, this paper provides the following results:

- (i) We demonstrate in Section 2 that the methodology introduced by Rissanen can be applied in a more general framework, and not only for the ellipsoidal constraints which have been considered in [18,25]. In the same section, we study the particular case of rhomboidal constraint.
- (ii) In Section 3, we conduct a rigorous analysis of the relationship between Rissanen criterion and the two criteria that have been introduced in [18].
- (iii) Section 4 is devoted to numerical examples which compare the capabilities of the NML-based selection rules against other criteria. The experiments are performed with simulated data as well as real life data sets.

Conclusions are outlined in Section 5, where we also give some guidance on the use of various criteria in model selection.

2. Parametric complexity with constraints

2.1. General case

To simplify the notations, we drop the index γ when discussing the general case. Let us define $\mathcal{Y}_\rho(R, \tau_0) = \{\mathbf{y} : \rho(\hat{\boldsymbol{\beta}}) \leq R, \hat{\tau} \geq \tau_0\}$, where R and τ_0 are strictly positive. The mapping $\rho : \mathbb{R}^k \rightarrow \mathbb{R}$ is chosen such that, for all $R > 0$, the set $\mathcal{B}_\rho(R) = \{\hat{\boldsymbol{\beta}} : \rho(\hat{\boldsymbol{\beta}}) \leq R\}$ is convex and its volume $V_\rho(R) = \int_{\mathcal{B}_\rho(R)} d\hat{\boldsymbol{\beta}}$ has the expression

$$V_\rho(R) = \eta R^{\zeta k}. \quad (10)$$

The constant η is strictly positive and, in some cases, it might depend on the regressor matrix \mathbf{X} . Additionally, the constant ζ is also assumed to be strictly positive.

Hence, the definition of NML from (8)–(9) is transformed to

$$\hat{f}_\rho(\mathbf{y}; R, \tau_0) = \frac{f(\mathbf{y}; \hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\tau}(\mathbf{y}))}{C_\rho(R, \tau_0)}, \quad (11)$$

$$C_\rho(R, \tau_0) = \int_{\mathcal{Y}_\rho(R, \tau_0)} f(\mathbf{y}; \hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\tau}(\mathbf{y})) d\mathbf{y}. \quad (12)$$

It is well known that the numerator in (11) is given by [31]

$$f(\mathbf{y}; \hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\tau}(\mathbf{y})) = [2\pi\hat{\tau}\exp(1)]^{-n/2}. \quad (13)$$

For the denominator, we prove in Appendix A that

$$C_\rho(R, \tau_0) = (2A_{n,k}/k)\tau_0^{-k/2}\eta R^{\zeta k}, \quad (14)$$

where

$$A_{n,k} = \frac{|\mathbf{X}^\top \mathbf{X}|^{1/2} \left(\frac{n}{2\exp(1)}\right)^{n/2}}{(n\pi)^{k/2} \Gamma\left(\frac{n-k}{2}\right)}. \quad (15)$$

The operator $|\cdot|$ denotes the determinant of the matrix in the argument.

Remark in (14) that the normalizing constant $C_\rho(R, \tau_0)$ becomes smaller when R decreases. Because we want to minimize the code length given by $-\ln \hat{f}_\rho(\mathbf{y}; R, \tau_0) = -\ln f(\mathbf{y}; \hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\tau}(\mathbf{y})) + \ln C_\rho(R, \tau_0)$, we assign to R the smallest possible value, namely $R = \hat{R}$, where $\hat{R} = \rho(\hat{\boldsymbol{\beta}})$. We choose $\hat{\tau}_0 = \hat{\tau}$ like in [26], and the expression from (11) becomes

$$\hat{f}_\rho(\mathbf{y}; \hat{R}, \hat{\tau}_0) = \frac{f(\mathbf{y}; \hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\tau}(\mathbf{y}))}{C_\rho(\hat{R}, \hat{\tau}_0)}. \quad (16)$$

Then we perform the second normalization step. Let $\mathcal{Y}(R_1, R_2, \tau_1, \tau_2) = \{\mathbf{y} : R_1 \leq \rho(\hat{\boldsymbol{\beta}}(\mathbf{y})) \leq R_2, \tau_1 \leq \hat{\tau}(\mathbf{y}) \leq \tau_2\}$, where $R_2 > R_1 > 0$ and $\tau_2 > \tau_1 > 0$. By using (16), we have

$$\hat{f}_\rho(\mathbf{y}) = \frac{\hat{f}_\rho(\mathbf{y}; \hat{R}, \hat{\tau}_0)}{\bar{C}_\rho(R_1, R_2, \tau_1, \tau_2)} = \frac{f(\mathbf{y}; \hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\tau}(\mathbf{y})) / C_\rho(\hat{R}, \hat{\tau}_0)}{\bar{C}_\rho(R_1, R_2, \tau_1, \tau_2)}. \quad (17)$$

The normalizing constant is given by

$$\bar{C}_\rho(R_1, R_2, \tau_1, \tau_2) = \int_{\mathcal{Y}(R_1, R_2, \tau_1, \tau_2)} \hat{f}_\rho(\mathbf{y}; \hat{R}, \hat{\tau}_0) d\mathbf{y}, \quad (18)$$

and after some calculations which are outlined in Appendix A, we obtain

$$\bar{C}_\rho(R_1, R_2, \tau_1, \tau_2) = \frac{\zeta k^2}{2} \ln \frac{\tau_2}{\tau_1} \ln \frac{R_2}{R_1}. \quad (19)$$

We collect the results from (13), (14), (17) and (19) to get the expression of the negative logarithm of NML, when the mapping $\rho(\cdot)$ is used to define the constraint for the evaluation of the parametric complexity:

$$\begin{aligned} -\ln \hat{f}_\rho(\mathbf{y}) &= -\ln f(\mathbf{y}; \hat{\boldsymbol{\beta}}, \hat{\tau}) + \ln C_\rho(\rho(\hat{\boldsymbol{\beta}}), \hat{\tau}) + \ln \bar{C}_\rho(R_1, R_2, \tau_1, \tau_2) \\ &= \frac{n-k}{2} \ln \hat{\tau} + \zeta k \ln \rho(\hat{\boldsymbol{\beta}}) - \ln \Gamma\left(\frac{n-k}{2}\right) + \ln \left[\zeta k \frac{\eta |\mathbf{X}^\top \mathbf{X}|^{1/2}}{(n\pi)^{k/2}} \right] \\ &\quad + \frac{n}{2} \ln(n\pi) + \ln \left(\ln \frac{\tau_2}{\tau_1} \ln \frac{R_2}{R_1} \right). \end{aligned} \quad (20)$$

It is obvious that, in the equations above, we have $\hat{\tau} = \hat{\tau}_\gamma$, $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_\gamma$ and $\mathbf{X} = \mathbf{X}_\gamma$. Conventionally we take $\tau_1 = R_1 = a$ and $\tau_2 = R_2 = b$, where $b > a > 0$. So,

$$-\ln \hat{f}_\rho(\mathbf{y}) = L_\rho(\mathbf{y}; \gamma) + L(a, b) + \frac{n}{2} \ln(n\pi), \quad (21)$$

$$L_\rho(\mathbf{y}; \gamma) = \frac{n-k}{2} \ln \hat{\tau}_\gamma + \zeta k \ln \rho(\hat{\boldsymbol{\beta}}_\gamma) - \ln \Gamma\left(\frac{n-k}{2}\right) + \ln \left[\zeta k \frac{\eta |\mathbf{X}_\gamma^\top \mathbf{X}_\gamma|^{1/2}}{(n\pi)^{k/2}} \right], \quad (22)$$

where $L(a, b)$ is the same as in (5).

For the sake of completeness, we consider also an approximate formula for the negative logarithm of NML [24]:

$$-\ln \hat{f}(\mathbf{y}) = -\ln f(\mathbf{y}; \hat{\boldsymbol{\beta}}, \hat{\tau}) + \frac{k+1}{2} \ln \frac{n}{2\pi} + \ln \int \mathbb{J}_\infty(\boldsymbol{\beta}, \tau)^{1/2} d\boldsymbol{\beta} d\tau + o(1), \quad (23)$$

where

$$\mathbb{J}_\infty(\boldsymbol{\beta}, \tau) = \lim_{n \rightarrow \infty} \mathbb{J}_n(\boldsymbol{\beta}, \tau), \quad (24)$$

$$\mathbb{J}_n(\boldsymbol{\beta}, \tau) = \begin{bmatrix} (\mathbf{X}^\top \mathbf{X})/(n\tau) & \mathbf{0} \\ \mathbf{0} & 1/(2\tau^2) \end{bmatrix}. \quad (25)$$

Remark in (23)–(25) that we have dropped the index γ . In (25), we have used the expression (see, for example, [14]) of the Fisher information matrix (FIM) for the linear model in (2). Note that, for many models used in signal processing, the right-hand side of (24) has a finite limit [36, Appendix C]. On contrary, the value of the integral in (23) is not finite if the domain of integration is the entire parameter space. This problem is well known and some of the proposed solutions involve arbitrarily chosen restrictions for the ranges of the parameters. A comprehensive discussion on this issue can be found in [11]. We demonstrate in Appendix A how the difficulty can be circumvented by applying constraints similar with those employed to get (20).

2.2. Rissanen formula

The constraint used by Rissanen is $\rho_1(\hat{\boldsymbol{\beta}}_\gamma) \leq R$, where $\rho_1(\hat{\boldsymbol{\beta}}_\gamma) = \|\mathbf{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma\|^2/n$ [26]. This makes the volume $V_{\rho_1}(R)$ to be given by (10) with $\eta = (n\pi)^{k/2}/[(k/2)\Gamma(k/2)|\mathbf{X}_\gamma^\top \mathbf{X}_\gamma|^{1/2}]$ and $\zeta = 1/2$. It is a simple exercise to show that, for the particular case when $\rho(\hat{\boldsymbol{\beta}}_\gamma) = \rho_1(\hat{\boldsymbol{\beta}}_\gamma)$, the formula

in (21)–(22) is identical with the one from (3)–(5). The expression of SC can be further simplified by operating the following modifications: (i) neglect the constant term $(n/2)\ln(n\pi)$ and the term $L(a, b)$; (ii) use the Stirling approximation (see Appendix A and [26,27])

$$\ln \Gamma(z) = (z - \frac{1}{2}) \ln z - z + \frac{1}{2} \ln(2\pi), \quad (26)$$

and then discard all terms which do not depend on the γ -structure; (iii) multiply by two the resulting criterion. This leads to

$$SC_{\rho_1}(\mathbf{y}; \gamma) = (n-k) \ln \frac{\hat{\tau}_\gamma}{n-k} + k \ln \frac{\|\mathbf{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma\|^2/n}{k} + \ln[k(n-k)]. \quad (27)$$

The above form of SC is the one which appeared most frequently in the literature after it was introduced in [25].

2.3. Rhomboidal constraint

Consider the constraint $\rho_0(\hat{\boldsymbol{\beta}}_\gamma) \leq R$, where $\rho_0(\hat{\boldsymbol{\beta}}_\gamma)$ is given by the 1-norm of $\hat{\boldsymbol{\beta}}_\gamma$, and we write $\rho_0(\hat{\boldsymbol{\beta}}_\gamma) = \|\hat{\boldsymbol{\beta}}_\gamma\|_1$. The region defined by the constraint is a diamond when $k=2$, and it becomes a rhomboid when $k > 2$ [12]. The volume $V_{\rho_0}(R)$ can be computed by observing that

$$V_{\rho_0}(R) = 2^k \times \int_{\substack{\hat{\beta}_1, \dots, \hat{\beta}_k \geq 0 \\ \hat{\beta}_1 + \dots + \hat{\beta}_k \leq R}} d\hat{\boldsymbol{\beta}}$$

because of the symmetry. Then we get $V_{\rho_0}(R) = (2R)^k/k!$. The result is easily verified for $k \in \{1, 2\}$ and is proven for any $k > 2$ by mathematical induction. More importantly, the formula which gives the volume $V_{\rho_0}(R)$ can be obtained from the one in (10) by choosing $\eta = 2^k/k!$ and $\zeta = 1$. Hence, we can get a new NML-based criterion by using in (21)–(22) the definition of $\rho_0(\cdot)$. For writing more compactly the new selection rule, we multiply by two the expression in (21), and we ignore the sum $2L(a, b) + n \ln(n\pi)$. Some elementary calculations lead to

$$\begin{aligned} SC_{\rho_0}(\mathbf{y}; \gamma) &= (n-k) \ln \hat{\tau}_\gamma + k \ln \frac{\|\hat{\boldsymbol{\beta}}_\gamma\|_1^2}{n} - 2 \ln \Gamma\left(\frac{n-k}{2}\right) - 2 \ln \Gamma(k) \\ &\quad + \ln \frac{4^k |\mathbf{X}_\gamma^\top \mathbf{X}_\gamma|}{\pi^k}. \end{aligned}$$

We modify the formula above by applying the Stirling approximation from (26), and by discarding the sum $(n-2)\ln 2 + n - 2\ln(2\pi)$, which was also neglected in (27). Thus, we have

$$\begin{aligned} SC_{\rho_0}(\mathbf{y}; \gamma) &= (n-k) \ln \frac{\hat{\tau}_\gamma}{n-k} + k \ln \frac{\|\hat{\boldsymbol{\beta}}_\gamma\|_1^2/n}{k} + \ln[k(n-k)] \\ &\quad + k \ln \frac{2 \exp(1)}{\pi k} + \ln(2|\mathbf{X}_\gamma^\top \mathbf{X}_\gamma|). \end{aligned} \quad (28)$$

From (27) and (28), it is obvious that the goodness-of-fit term is the same for both $SC_{\rho_1}(\mathbf{y}; \gamma)$ and $SC_{\rho_0}(\mathbf{y}; \gamma)$. We want to check which is the relationship between the penalty terms of the two criteria. For ease of comparison, we assume that the columns of \mathbf{X}_γ are the first k columns of the $n \times n$ identity matrix, which implies

$$\text{PEN}_{\rho_0}(\mathbf{y}; \gamma) - \text{PEN}_{\rho_1}(\mathbf{y}; \gamma) = k \ln \frac{2 \exp(1) \cos^2(\alpha_\gamma)}{\pi} + \ln 2, \quad (29)$$

where $\cos \alpha_\gamma = \|\hat{\beta}_\gamma\|_1 / (\sqrt{k} \|\hat{\beta}_\gamma\|)$. Equivalently, α_γ is the angle between the vector $[\|\hat{\beta}_1\|, \dots, \|\hat{\beta}_k\|]^\top$, which is given by the magnitudes of the estimates, and the vector $[1, \dots, 1]^\top$. Remark in (29) that $\text{PEN}_{\rho_0}(\mathbf{y}; \gamma) - \text{PEN}_{\rho_1}(\mathbf{y}; \gamma) > 0$ if and only if $\alpha_\gamma < \arccos(\text{Th}_k)$, where $\text{Th}_k = \{\pi / (2\exp(1))\}^{1/2}$. For all $k \geq 1$, the inequality $\text{Th}_k \in (0, 1)$ is satisfied, and for $k \gg 1$, we have $\arccos(\text{Th}_k) \approx (2\pi)/9$.

To gain more insight, we assume that $\mathbf{y} \sim \mathcal{N}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\tau}}\mathbf{I})$, where $\mathcal{N}(\boldsymbol{\mu}, \mathbf{R})$ denotes the multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix \mathbf{R} . The vector $\bar{\boldsymbol{\mu}}$ is chosen such that to have \bar{k} entries equal to a non-zero constant β , and all other entries are zeros. Additionally, $\bar{k} \ll n$, and the value of $\bar{\tau}$ is selected to guarantee a certain signal-to-noise ratio. Let $y_{(1)}, \dots, y_{(n)}$ be the measurements sorted in the decreasing order of their magnitudes. For each $k \in \{1, \dots, n-2\}$, we define the structure $\gamma_k = \{(1), \dots, (k)\}$ such that $\hat{\beta}_{\gamma_k} = [y_{(1)}, \dots, y_{(k)}]^\top$ and $\hat{\tau}_{\gamma_k} = (1/n) \sum_{i=k+1}^n y_{(i)}^2$. When $k > \bar{k}$, if k increases, then the angle α_{γ_k} increases also, and $\text{PEN}_{\rho_0}(\mathbf{y}; \gamma_k)$ becomes smaller than $\text{PEN}_{\rho_1}(\mathbf{y}; \gamma_k)$. Hence, the criterion SC_{ρ_0} penalizes less than SC_{ρ_1} when k is large, which makes to be more likely that SC_{ρ_1} selects a sparser solution, and not SC_{ρ_0} .

This outcome is surprising because it is known from the previous literature [12, Chapter 3] that the selection rules which have as penalty term the 1-norm of the vector of estimates are prone to pick-up the sparse solutions.

The formulas derived with the general methodology described in Section 2.1 must be used with caution in practice, and only after their properties are carefully investigated. Next, we focus on two other NML-based criteria, which have been introduced in [18] to cope with the presence of collinearity.

3. Ellipsoidal constraint

3.1. Formulas from [18]

The solution proposed in [18] for the computation of the parametric complexity relies on the following ellipsoidal constraint: $(\hat{\beta}_\gamma^\top \mathbf{Q} \hat{\beta}_\gamma) / n \leq R$, where the matrix \mathbf{Q} is chosen to be symmetric and positive definite. By applying the formula for the volume of an ellipsoid [30], it is easy to verify for $\rho(\hat{\beta}_\gamma) = \hat{\beta}_\gamma^\top \mathbf{Q} \hat{\beta}_\gamma$ that $V_\rho(R)$ is a particular case of (10) for which $\eta = (n\pi)^{k/2} / [(k/2)\Gamma(k/2)|\mathbf{Q}|^{1/2}]$ and $\zeta = 1/2$. By employing in (21)–(22) the expressions of η and ζ , we have

$$-\ln \hat{f}_\rho(\mathbf{y}) = \frac{n-k}{2} \ln \hat{\tau}_\gamma + \frac{k}{2} \ln \frac{\hat{\beta}_\gamma^\top \mathbf{Q} \hat{\beta}_\gamma}{n} - \ln \Gamma\left(\frac{n-k}{2}\right) - \ln \Gamma\left(\frac{k}{2}\right) + \frac{1}{2} \ln \frac{|\mathbf{X}_\gamma^\top \mathbf{X}_\gamma|}{|\mathbf{Q}|}, \quad (30)$$

which coincides with [18, Eq. (16)]. Remark in (30) that we have neglected the terms $L(a, b)$ and $(n/2) \ln(n\pi)$.

When $\mathbf{Q} = \mathbf{X}_\gamma^\top \mathbf{X}_\gamma$, the ellipsoidal constraint $(\hat{\beta}_\gamma^\top \mathbf{Q} \hat{\beta}_\gamma) / n \leq R$ is identical with the constraint used by Rissanen, namely $\rho_1(\hat{\beta}_\gamma) \leq R$. Other two possible ways of selecting the matrix \mathbf{Q} have been considered in [18]: $\mathbf{Q} = \mathbf{I}$ and $\mathbf{Q} = (\mathbf{X}_\gamma^\top \mathbf{X}_\gamma)^2$. In the case when $\mathbf{Q} = \mathbf{I}$, the ellipsoidal constraint becomes $\rho_2(\hat{\beta}_\gamma) \leq R$, where $\rho_2(\hat{\beta}_\gamma) = \|\hat{\beta}_\gamma\|^2 / n$. By operating

in (30) the same type of modifications which allowed to transform (3)–(5) into (27), we get

$$\text{SC}_{\rho_2}(\mathbf{y}; \gamma) = (n-k) \ln \frac{\hat{\tau}_\gamma}{n-k} + k \ln \frac{\|\hat{\beta}_\gamma\|^2 / n}{k} + \ln[k(n-k)] + \ln |\mathbf{X}_\gamma^\top \mathbf{X}_\gamma|. \quad (31)$$

Similarly, for $\mathbf{Q} = (\mathbf{X}_\gamma^\top \mathbf{X}_\gamma)^2$, the ellipsoidal constraint takes the form $\rho_3(\hat{\beta}_\gamma) \leq R$ with $\rho_3(\hat{\beta}_\gamma) = [\hat{\beta}_\gamma^\top (\mathbf{X}_\gamma^\top \mathbf{X}_\gamma)^2 \hat{\beta}_\gamma] / n$, and the corresponding model selection criterion is

$$\text{SC}_{\rho_3}(\mathbf{y}; \gamma) = (n-k) \ln \frac{\hat{\tau}_\gamma}{n-k} + k \ln \frac{[\hat{\beta}_\gamma^\top (\mathbf{X}_\gamma^\top \mathbf{X}_\gamma)^2 \hat{\beta}_\gamma] / n}{k} + \ln[k(n-k)] - \ln |\mathbf{X}_\gamma^\top \mathbf{X}_\gamma|. \quad (32)$$

After discarding the term $-n \ln n$ from the formulas in (27), (31) and (32), we can re-write them as follows. For $i \in \{1, 2, 3\}$,

$$\text{SC}_{\rho_i}(\mathbf{y}; \gamma) = (n-k) \ln S_\gamma^2 + k \ln D_\gamma(\mathbf{y}; \mathbf{Q}_i) + \ln \frac{n-k}{k^{k-1}}, \quad (33)$$

$$D_\gamma(\mathbf{y}; \mathbf{Q}_i) = \frac{\mathbf{y}^\top \mathbf{X}_\gamma (\mathbf{X}_\gamma^\top \mathbf{X}_\gamma)^{-1} \mathbf{Q}_i (\mathbf{X}_\gamma^\top \mathbf{X}_\gamma)^{-1} \mathbf{X}_\gamma^\top \mathbf{y}}{(|\mathbf{Q}_i| / |\mathbf{X}_\gamma^\top \mathbf{X}_\gamma|)^{1/k}}, \quad (34)$$

where $S_\gamma^2 = (n\hat{\tau}_\gamma) / (n-k)$, $\mathbf{Q}_1 = \mathbf{X}_\gamma^\top \mathbf{X}_\gamma$, $\mathbf{Q}_2 = \mathbf{I}$ and $\mathbf{Q}_3 = (\mathbf{X}_\gamma^\top \mathbf{X}_\gamma)^2$. It is evident that all three selection rules have the same goodness-of-fit term, and only $D_\gamma(\mathbf{y}; \mathbf{Q}_i)$ makes their penalty terms to be different.

3.2. Penalty terms

For better understanding the relationship between the three criteria, we give the following result.

Proposition 3.1.

(a) The equalities

$$D_\gamma(\mathbf{y}; \mathbf{Q}_1) = D_\gamma(\mathbf{y}; \mathbf{Q}_2) = D_\gamma(\mathbf{y}; \mathbf{Q}_3) \quad (35)$$

hold true for all $\mathbf{y} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ if and only if there exists $q > 0$ such that

$$\mathbf{Q}_1 = q\mathbf{I}. \quad (36)$$

(b) If the condition in (36) is not satisfied, then for each pair (i, j) with the property that $1 \leq i < j \leq 3$, the sign of the difference

$$D_\gamma(\mathbf{y}; \mathbf{Q}_i) - D_\gamma(\mathbf{y}; \mathbf{Q}_j)$$

is not the same for all $\mathbf{y} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$.

(c) For all $\mathbf{y} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$, we have

$$\max\{D_\gamma(\mathbf{y}; \mathbf{Q}_2), D_\gamma(\mathbf{y}; \mathbf{Q}_3)\} \geq D_\gamma(\mathbf{y}; \mathbf{Q}_1). \quad (37)$$

Proof is deferred to Appendix B.

From the proposition above, we see that the criteria $\text{SC}_{\rho_1}(\mathbf{y}; \gamma)$, $\text{SC}_{\rho_2}(\mathbf{y}; \gamma)$ and $\text{SC}_{\rho_3}(\mathbf{y}; \gamma)$ are identical only when the columns of the matrix \mathbf{X}_γ are orthogonal and the 2-norm is the same for all of them. In general, it is not possible to claim that one criterion has a penalty term which is stronger than the penalty terms of the others. However, the inequality in (37) guarantees that at least one of the criteria $\text{SC}_{\rho_2}(\mathbf{y}; \gamma)$ and $\text{SC}_{\rho_3}(\mathbf{y}; \gamma)$ has a penalty term which is stronger than the penalty term of the Rissanen criterion.

Next we investigate the behavior of the three selection rules for the case when the matrix \mathbf{X}_γ is rank deficient. Let us use the notation \mathbf{X}_k instead of \mathbf{X}_γ . Furthermore, we partition the matrix into two blocks: $\mathbf{X}_k = [\mathbf{X}_{k-1} \ \mathbf{x}_k]$. Note that \mathbf{X}_{k-1} contains the first $k-1$ columns of \mathbf{X}_k . We assume that \mathbf{X}_{k-1} has full-rank, and the source of rank deficiency for \mathbf{X}_k is the fact that the linear subspaces $\langle \mathbf{X}_{k-1} \rangle$ and $\langle \mathbf{x}_k \rangle$ are “very close” to each other. For a full-rank matrix \mathbf{M} having more rows than columns, $\langle \mathbf{M} \rangle$ is the column space of \mathbf{M} .

To measure the “closeness”, we employ the *principal angle* $\alpha \in [0, \pi/2]$ between $\langle \mathbf{X}_{k-1} \rangle$ and $\langle \mathbf{x}_k \rangle$ [3]. If the columns of \mathbf{U}_{k-1} form a unitary basis for $\langle \mathbf{X}_{k-1} \rangle$ and \mathbf{u}_k is a unitary basis for $\langle \mathbf{x}_k \rangle$, then $\cos \alpha$ is the singular value of $\mathbf{U}_{k-1}^\top \mathbf{u}_k$. Eq. (13) from [3] guarantees that there exists $\mathbf{w} \in \mathbb{R}^{n \times 1}$ with $\|\mathbf{w}\| = 1$ such that

$$\mathbf{P}_{k-1}^\perp \mathbf{x}_k = \sin(\alpha) \|\mathbf{x}_k\| \mathbf{w}, \quad (38)$$

where $\mathbf{P}_{k-1}^\perp = \mathbf{I} - \mathbf{P}_{k-1}$ and $\mathbf{P}_{k-1} = \mathbf{X}_{k-1} (\mathbf{X}_{k-1}^\top \mathbf{X}_{k-1})^{-1} \mathbf{X}_{k-1}^\top$ is the orthogonal projection matrix onto the linear subspace $\langle \mathbf{X}_{k-1} \rangle$. The following proposition clarifies which is the effect of $\alpha \rightarrow 0$ on the penalty terms.

Proposition 3.2. *If $\text{rank}(\mathbf{X}_{k-1}) = k-1$, $\|\mathbf{x}_k\| \neq 0$, $k > 1$ and $\mathbf{y} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$, then:*

- (a) $\lim_{\alpha \rightarrow 0} D_\gamma(\mathbf{y}; \mathbf{Q}_1) < \infty$.
- (b) $\lim_{\alpha \rightarrow 0} D_\gamma(\mathbf{y}; \mathbf{Q}_2) = \infty$ when $\mathbf{w}^\top \mathbf{y} \neq 0$. Note that \mathbf{w} is defined in (38).
- (c) $\lim_{\alpha \rightarrow 0} D_\gamma(\mathbf{y}; \mathbf{Q}_3) = \infty$ when $\mathbf{X}_k^\top \mathbf{y} \neq \mathbf{0}$.

See Appendix B for the proof.

Remark that, under the assumptions from Proposition 3.2, SC_{ρ_2} and SC_{ρ_3} penalize the collinearity more severely than SC_{ρ_1} . The result has to be understood in connection with the fact that variable selection aims to discard those columns of \mathbf{X} which are nearly collinear, and then to use the retained columns for explaining the variation in \mathbf{y} [19, Section 6.7]. This can be nicely formalized by using the *coefficient of determinations* whose definitions are given below.

Definition 3.1. Assume that the sum of the entries of \mathbf{y} is zero and $\|\mathbf{y}\| = 1$. Additionally, each column of \mathbf{X} is zero-mean and has unitary Euclidean norm. For an arbitrary γ -structure with cardinality $k > 0$, we define

$$R_{\mathbf{y}, \mathbf{X}_\gamma}^2 = \|\mathbf{P}_\gamma \mathbf{y}\|^2, \quad (39)$$

where \mathbf{P}_γ is the orthogonal projection matrix onto the linear subspace $\langle \mathbf{X}_\gamma \rangle$. Moreover, for $i \in \{2, \dots, k\}$, we have

$$R_{i,1,\dots,(i-1)}^2 = \|\mathbf{P}_{i-1} \mathbf{x}_i\|^2, \quad (40)$$

where \mathbf{P}_{i-1} denotes the orthogonal projection matrix onto the linear subspace determined by the first $(i-1)$ columns of \mathbf{X}_γ , and \mathbf{x}_i is the i -th column of \mathbf{X}_γ .

It is clear that (39) and (40) are just particular cases of the general definition that can be found in [19, Section 6.5.2; 31, p. 111]. We emphasize that $R_{\mathbf{y}, \mathbf{X}_\gamma}^2$ is a measure of how much the variance of $\mathbf{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma$ represents from the total variance of the data \mathbf{y} . A similar interpretation can be given for (40).

In the next proposition, we show how the dependence of \mathbf{y} on \mathbf{X}_γ , as well as the interdependence between the columns of \mathbf{X}_γ , affect the terms $D_\gamma(\mathbf{y}; \mathbf{Q}_1)$, $D_\gamma(\mathbf{y}; \mathbf{Q}_2)$ and $D_\gamma(\mathbf{y}; \mathbf{Q}_3)$.

Proposition 3.3. *When \mathbf{X}_γ and \mathbf{y} satisfy the conditions from Definition 3.1, the following identities hold true:*

$$D_\gamma(\mathbf{y}; \mathbf{Q}_1) = R_{\mathbf{y}, \mathbf{X}_\gamma}^2, \quad (41)$$

$$D_\gamma(\mathbf{y}; \mathbf{Q}_2) = \sum_{i=1}^k a_i(\mathbf{y}, \mathbf{X}_\gamma) b_i(\mathbf{X}_\gamma), \quad (42)$$

$$D_\gamma(\mathbf{y}; \mathbf{Q}_3) = \frac{\sum_{i=1}^k r_{i\mathbf{y}}^2}{\prod_{i=2}^k [1 - R_{i,1,\dots,(i-1)}^2]^{1/k}}, \quad (43)$$

where

$$a_i(\mathbf{y}, \mathbf{X}_\gamma) = R_{\mathbf{y}, \mathbf{X}_\gamma}^2 - R_{\mathbf{y}, \mathbf{X}_{\gamma(i)}}^2,$$

$$b_i(\mathbf{X}_\gamma) = \frac{\prod_{j=2}^{k-1} [1 - R_{\zeta(j), \zeta(1), \dots, \zeta(j-1)}^2]^{1/k}}{[1 - R_{\zeta(k), \zeta(1), \dots, \zeta(k-1)}^2]^{(k-1)/k}},$$

$$\zeta(j) = \begin{cases} j, & 1 \leq j < i, \\ j+1, & i \leq j < k, \\ i, & j = k, \end{cases} \quad (44)$$

and $r_{i\mathbf{y}}$ is the correlation between the i -th column of \mathbf{X}_γ and \mathbf{y} .

See Appendix B for the proof.

Eq. (41) confirms that the interdependence between the columns of \mathbf{X}_γ does not have any impact on $D_\gamma(\mathbf{y}; \mathbf{Q}_1)$. This is not the case with $D_\gamma(\mathbf{y}; \mathbf{Q}_2)$, where the factors $b_i(\mathbf{X}_\gamma)$ measure the linear dependence between the columns of \mathbf{X}_γ , and they are not affected by the relationship between \mathbf{y} and \mathbf{X}_γ . Whenever \mathbf{x}_i is a linear combination of some of other columns from \mathbf{X}_γ , the denominator of $b_i(\mathbf{X}_\gamma)$ becomes zero, whereas the numerator is strictly positive. In this situation, the contribution of \mathbf{x}_i to explaining the variance of \mathbf{y} is marginal, which makes $a_i(\mathbf{y}, \mathbf{X}_\gamma)$ to be also zero. From (43), it is evident how multicollinearity affects $D_\gamma(\mathbf{y}; \mathbf{Q}_3)$: the denominator goes to zero and the nominator remains strictly positive.

Propositions 3.1 and 3.2 reveal the relationship between the three criteria when the angle α takes extreme values: $\alpha = \pi/2$ and $\alpha \rightarrow 0$. It remains open the question on how SC_{ρ_2} and SC_{ρ_3} relate to SC_{ρ_1} when $\alpha \in (0, \pi/2)$. In order to answer the question, we need to make supplementary assumptions on the vector of observations \mathbf{y} . This is why we consider next the case of two nested models.

3.3. Comparison of the penalty terms when two nested models are tested

Suppose that the model selection problem reduces to deciding if the measurements \mathbf{y} are outcomes from $\mathcal{N}(\mathbf{X}_{k-1} \boldsymbol{\beta}_{k-1}, \tau \mathbf{I})$ or from $\mathcal{N}(\mathbf{X}_k \boldsymbol{\beta}_k, \tau \mathbf{I})$, where \mathbf{X}_{k-1} and \mathbf{X}_k are the same as in Proposition 3.2. The entries of $\boldsymbol{\beta}_{k-1} \in \mathbb{R}^{k-1}$ and $\boldsymbol{\beta}_k \in \mathbb{R}^k$ are assumed to be non-zero, and $\tau > 0$. After estimating $\hat{\boldsymbol{\beta}}_{k-1}$, $\hat{\boldsymbol{\beta}}_k$ and the noise variance from the available data, one can apply an NML-based criterion to

select between the structure γ_{k-1} for which the regression matrix is \mathbf{X}_{k-1} , and the structure γ_k for which the regression matrix is \mathbf{X}_k .

We know from Proposition 3.1 that, disregarding the machinery which has produced \mathbf{y} , we have $SC_{\rho_1}(\mathbf{y}; \gamma_{k-1}) = SC_{\rho_2}(\mathbf{y}; \gamma_{k-1}) = SC_{\rho_3}(\mathbf{y}; \gamma_{k-1})$ if $\mathbf{X}_{k-1}^\top \mathbf{X}_{k-1} = \mathbf{I}$. Therefore, under the hypothesis of orthonormality for the columns of \mathbf{X}_{k-1} , $D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_i)$, $i \in \{1, 2, 3\}$, is the only term which can potentially make SC_{ρ_1} , SC_{ρ_2} , SC_{ρ_3} not to take the same decision when choosing between γ_{k-1} and γ_k . To gain more insight, we compute the expectation of $D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_i)$ for $i \in \{1, 2, 3\}$.

Lemma 3.1. *If $\mathbf{y} \sim \mathcal{N}(\mathbf{X}_{k-1}\boldsymbol{\beta}_{k-1}, \tau\mathbf{I})$, $\mathbf{X}_{k-1}^\top \mathbf{X}_{k-1} = \mathbf{I}$, $\|\mathbf{x}_k\| = 1$ and $k > 1$, then*

$$\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)] = \|\boldsymbol{\beta}_{k-1}\|^2 + \tau k, \quad (45)$$

$$\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_2)] = [\|\boldsymbol{\beta}_{k-1}\|^2 + \tau(k-2+2\omega^{-1})]\omega^{1/k}, \quad (46)$$

$$\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_3)] = [\|\boldsymbol{\beta}_{k-1}\|^2 + \tau k + (\mathbf{x}_k^\top \mathbf{X}_{k-1} \boldsymbol{\beta}_{k-1})^2]\omega^{-1/k}, \quad (47)$$

where $\mathbb{E}[\cdot]$ is the expectation operator and $\omega = \sin^2 \alpha$.

The proof of Lemma 3.1 can be found in Appendix C, where we outline also the proof of the proposition below.

Proposition 3.4. *Let $\varphi_0(\alpha) = 2(1 - \sin^2 \alpha)/(1 - \sin^{2/k} \alpha) - k$, $\varphi_1(\alpha) = (2 - \sin^2 \alpha)\sin^{-2/k} \alpha - 1$ and $\varphi_2(\alpha) = \sin^{-2/k} \alpha - 1$. Under the hypotheses of Lemma 3.1, we have:*

- (a) $\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_2) - D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)]$ is strictly positive if and only if $\alpha < \alpha^*$, where α^* is the solution of the equation $\varphi_0(\alpha) = \|\boldsymbol{\beta}_{k-1}\|^2 / \tau$.
- (b) $\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_3) - D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)]$ takes only non-negative values. Additionally,

$$\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_3) - D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)] \leq \|\boldsymbol{\beta}_{k-1}\|^2 \varphi_1(\alpha) + \tau k \varphi_2(\alpha). \quad (48)$$

In Proposition 3.4, the Rissanen formula (27) is considered to be a reference, and the other two criteria are compared with it. We see that SC_{ρ_2} is likely to penalize more than SC_{ρ_1} the model with structure γ_k only when the angle between $\langle \mathbf{X}_{k-1} \rangle$ and $\langle \mathbf{x}_k \rangle$ is smaller than a threshold. The value of the threshold is mainly given by the ratio $\|\boldsymbol{\beta}_{k-1}\|^2 / \tau$, which in our case equals the energy-to-noise ratio (ENR) because $\|\mathbf{X}_{k-1} \boldsymbol{\beta}_{k-1}\| = \|\boldsymbol{\beta}_{k-1}\|$. Since $\lim_{\alpha \rightarrow 0} \varphi_0(\alpha) = \infty$ and $\lim_{\alpha \rightarrow \pi/2} \varphi_0(\alpha) = k$, the solution α^* is guaranteed to exist when $\text{ENR} > k$. Moreover, if ENR is larger than k , then $\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_2) - D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)]$ attains its minimum when the principal angle takes value $\alpha_{\min} = \arcsin(\omega_{\min}^{1/2})$, where $\omega_{\min} = 2(k-1)/(\|\boldsymbol{\beta}_{k-1}\|^2 / \tau + (k-2))$. The increase of ENR makes α_{\min} to decrease and α^* to be closer to zero such that SC_{ρ_2} penalizes more severely than SC_{ρ_1} only when $\alpha \approx 0$.

On contrary, SC_{ρ_3} penalizes the γ_k -model more stringently than SC_{ρ_1} for all $\alpha \in (0, \pi/2)$. Observe in (48) that $\varphi_1(\alpha)$ and $\varphi_2(\alpha)$ are monotonically decreasing functions, and the upper bound for $\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_3) - D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)]$ goes down from ∞ to zero when α increases from zero to $\pi/2$.

To complete the analysis, we provide the analogue of Proposition 3.4 for the case when $\mathbf{y} \sim \mathcal{N}(\mathbf{X}_k \boldsymbol{\beta}_k, \tau \mathbf{I})$. Let us assume that the eigenvalues of $\mathbf{X}_k^\top \mathbf{X}_k$ are $\lambda_1, \dots, \lambda_k$, and all of them are strictly positive. To write more compactly the

results, we define: $\mathcal{A}_\lambda = (\sum_{i=1}^k \lambda_i)/k$ (arithmetic mean), $\mathcal{G}_\lambda = (\prod_{i=1}^k \lambda_i)^{1/k}$ (geometric mean) and $\mathcal{H}_\lambda = k / \sum_{i=1}^k \lambda_i^{-1}$ (harmonic mean). The expressions of $\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_i)]$ for $i \in \{1, 2, 3\}$ are given in the lemma below.

Lemma 3.2. *If $\mathbf{y} \sim \mathcal{N}(\mathbf{X}_k \boldsymbol{\beta}_k, \tau \mathbf{I})$, $\mathbf{X}_{k-1}^\top \mathbf{X}_{k-1} = \mathbf{I}$, $\|\mathbf{x}_k\| = 1$ and $k > 1$, then*

$$\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)] = \|\mathbf{X}_k \boldsymbol{\beta}_k\|^2 + \tau k, \quad (49)$$

$$\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_2)] = \|\boldsymbol{\beta}_k\|^2 \mathcal{G}_\lambda + \tau k (\mathcal{G}_\lambda / \mathcal{H}_\lambda), \quad (50)$$

$$\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_3)] = [\boldsymbol{\beta}_k^\top (\mathbf{X}_k^\top \mathbf{X}_k)^2 \boldsymbol{\beta}_k] / \mathcal{G}_\lambda + \tau k (\mathcal{A}_\lambda / \mathcal{G}_\lambda). \quad (51)$$

Proof. The results are easily obtained by applying the formula of the expectation for quadratic forms [30, p. 439]. \square

Lemma 3.2 helps us to find bounds for $\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_2) - D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)]$ and $\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_3) - D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)]$, which are similar with those given in Proposition 3.4.

Proposition 3.5. *Under the hypotheses of Lemma 3.2, we have*

- (a) Let $\psi_1(\alpha) = \sin^{2/k} \alpha - \cos \alpha - 1$, $\psi_2(\alpha) = \sin^{2/k} \alpha + \cos \alpha - 1$ and $\psi_3(\alpha) = \sin^{2/k} \alpha / (1 - \cos \alpha) - 1$. Then

$$\begin{aligned} \|\boldsymbol{\beta}_k\|^2 \psi_1(\alpha) &\leq \mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_2) - D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)] \\ &\leq \|\boldsymbol{\beta}_k\|^2 \psi_2(\alpha) + \tau k \psi_3(\alpha). \end{aligned} \quad (52)$$

- (b) Let $\psi_4(\alpha) = -\sin^{2/k} \alpha / 4$, $\psi_5(\alpha) = (1 - \cos \alpha)^2 / \sin^{2/k} \alpha + \cos \alpha - 1$, $\psi_6(\alpha) = (1 + \cos \alpha)^2 / \sin^{2/k} \alpha - \cos \alpha - 1$ and $\psi_7(\alpha) = (1 + \cos \alpha) / \sin^{2/k} \alpha - 1$. For $\alpha \in (0, \pi/2]$, $\|\boldsymbol{\beta}_k\|^2 \psi_4(\alpha) \leq \mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_3) - D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)]$, (53)

and for $\alpha \in [\pi/3, \pi/2]$, the inequality becomes

$$\|\boldsymbol{\beta}_k\|^2 \psi_4(\alpha) \leq \|\boldsymbol{\beta}_k\|^2 \psi_5(\alpha) \leq \mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_3) - D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)]. \quad (54)$$

Additionally,

$$\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_3) - D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)] \leq \|\boldsymbol{\beta}_k\|^2 \psi_6(\alpha) + \tau k \psi_7(\alpha), \quad (55)$$

for all $\alpha \in (0, \pi/2]$.

Proof is deferred to Appendix C.

Note in (52) that the span of $\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_2) - D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)]$ is given by $\|\boldsymbol{\beta}_k\|^2 [\psi_2(\alpha) - \psi_1(\alpha)] + \tau k \psi_3(\alpha)$. The second term is the dominant one when α is close to zero, as we can see from $\lim_{\alpha \rightarrow 0} \{\|\boldsymbol{\beta}_k\|^2 [\psi_2(\alpha) - \psi_1(\alpha)]\} = 2\|\boldsymbol{\beta}_k\|^2 < \infty$ and $\lim_{\alpha \rightarrow 0} \{\tau k \psi_3(\alpha)\} = \infty$. To monitor the decrease of the two terms when α varies from zero to $\pi/2$, we define $\mathfrak{F}_{\psi_1, \psi_2}(\alpha_1, \alpha_2) = (\psi_2(\alpha_2) - \psi_1(\alpha_2)) / (\psi_2(\alpha_1) - \psi_1(\alpha_1))$ and $\mathfrak{F}_{\psi_3}(\alpha_1, \alpha_2) = \psi_3(\alpha_2) / \psi_3(\alpha_1)$, where $0 < \alpha_1 < \alpha_2 < \pi/2$. For example, when $k=6$, we get $\mathfrak{F}_{\psi_1, \psi_2}(\pi/180, \pi/6) \approx 87\%$, $\mathfrak{F}_{\psi_1, \psi_2}(\pi/6, \pi/3) \approx 58\%$ and $\mathfrak{F}_{\psi_1, \psi_2}(\pi/3, \pi/2 - \pi/180) \approx 3\%$, whereas $\mathfrak{F}_{\psi_3}(\pi/180, \pi/6) \approx 0.3\%$, $\mathfrak{F}_{\psi_3}(\pi/6, \pi/3) \approx 18\%$ and $\mathfrak{F}_{\psi_3}(\pi/3, \pi/2 - \pi/180) \approx 2\%$. Remark that the term given by $\psi_3(\cdot)$ diminishes significantly when α increases from $\pi/180$ to $\pi/6$. Another significant reduction occurs for both terms in the interval $[\pi/6, \pi/2 - \pi/180]$. An important observation is that the upper bound in (52) increases

monotonically with τ when \mathbf{X}_k and β_k are fixed. Therefore, when the ENR lowers, there exists a higher chance that SC_{ρ_2} penalizes the γ_k -model more stringently than SC_{ρ_1} . This finding is of special interest because, in Proposition 3.5, the model with structure γ_k is assumed to be the “true” one.

A similar analysis can be done for $\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_3) - D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)]$. In the vicinity of zero, the span of $\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_3) - D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)]$ is given by $\|\beta_k\|^2[\psi_6(\alpha) - \psi_4(\alpha)]$ and $\tau k \psi_7(\alpha)$. Since $\lim_{\alpha \rightarrow 0} \{\|\beta_k\|^2[\psi_6(\alpha) - \psi_4(\alpha)]\} = \lim_{\alpha \rightarrow 0} \{\tau k \psi_7(\alpha)\} = \infty$, the two terms are equally important, not like in the case of $\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_2) - D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)]$. It is also interesting to point out for $k=6$ that $(\psi_6(\pi/2 - \pi/180) - \psi_5(\pi/2 - \pi/180))/(\psi_6(\pi/3) - \psi_5(\pi/3)) \approx \psi_7(\pi/2 - \pi/180)/\psi_7(\pi/3) \approx 3\%$, which is similar with the result found previously for $\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_2) - D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)]$.

4. Experimental results

4.1. Model selection criteria used in experiments

We illustrate the performance of SC_{ρ_1} , SC_{ρ_2} and SC_{ρ_3} against other criteria. For ease of comparison, we employ for all the model selection rules the same notations like those from (3)–(4). As already told in Section 1, BIC is among the most popular criteria, and this is why we include it in our experiments. The well-known expression of BIC is [29]

$$\text{BIC}(\mathbf{y}; \gamma) = \frac{n}{2} \ln \hat{\tau}_\gamma + \frac{k}{2} \ln n. \quad (56)$$

Another widely used criteria are AIC and its bias corrected version which is called AIC_c [13]. Recently, Seghouane has applied bootstrap-type techniques to obtain AIC_{c3} , a new corrected version of AIC. The complete derivation can be found in [32], where it was also shown experimentally that, for small sample size, AIC_{c3} outperforms AIC_c as well as two other corrected criteria: AIC_c^* [33] and KIC_c [34]. Remark that the small sample size case makes the difference between various forms of AIC because asymptotically all of them are equivalent. For the sake of comparison, we consider in our simulations the criterion from [32]:

$$\text{AIC}_{c3}(\mathbf{y}; \gamma) = \frac{n}{2} \ln \hat{\tau}_\gamma + \frac{(k+1)(n+k+2)}{n-k-2} - \frac{k}{n-k}. \quad (57)$$

Following the suggestion of one of the reviewers, we briefly discuss how SC_{ρ_1} relates to BIC and AIC. The aim of the discussion is to provide support for the interpretation of the experimental results presented in this section. Note that the formula of SC_{ρ_1} from (27) can be re-written as follows [7, Eq. (16)]:

$$\frac{1}{2} SC_{\rho_1}(\mathbf{y}; \gamma) = \frac{n}{2} \ln \hat{\tau}_\gamma + \frac{k}{2} \ln F_\gamma + \frac{1}{2} \ln \frac{k}{(n-k)^{n-1}}, \quad (58)$$

where $F_\gamma = (\|\mathbf{X}_\gamma \hat{\beta}_\gamma\|^2 / (nk)) / (\hat{\tau}_\gamma / (n-k))$. It is evident that the goodness-of-fit term is the same for all the criteria in (56)–(58). The key difference is that F_γ from (58) depends on the data vector \mathbf{y} , while the penalty terms from (56) and (57) depend only on n and k . Let us observe that F_γ coincides with the F -statistic which is used to test the hypothesis that each entry of $\hat{\beta}_\gamma$ is zero [17, Section 5; 31, Chapter 4].

More importantly, by applying the settings from [4], it was worked out in [10] an expression of F_γ which leads to the conclusion that, asymptotically, SC_{ρ_1} combines the strengths of both BIC and AIC. Similarly with BIC, SC_{ρ_1} is *consistent*: if the “true model” is finite-dimensional and is included in the set of candidates, then the probability that this model is selected goes to one as the sample size increases [8]. However, if the “true model” is not finite-dimensional, then SC_{ρ_1} is asymptotically *efficient* in the sense that selects the candidate model which minimizes the one-step mean squared error of prediction. The same property has been proved for AIC long time ago [35]. We refer to [10] for the technical details concerning the results outlined above.

The two-part MDL criterion, which is equivalent to BIC, was refined in [22] such that its penalty term involves the logarithm of determinant of the observed FIM. A similar formula, which is not rooted in information theory, was proposed by Kay [15]:

$$\text{CME}(\mathbf{y}; \gamma) = \frac{n-k-2}{2} \ln \frac{n \hat{\tau}_\gamma}{n-k} + \frac{1}{2} \ln |\mathbf{X}_\gamma^\top \mathbf{X}_\gamma| + \ln \frac{[\pi(n-k)]^{(n-k)/2}}{\Gamma(\frac{n-k}{2})}. \quad (59)$$

The significance of the acronym CME is conditional model estimator.

In addition to BIC, AIC_{c3} and CME, we include in our tests the MML_g criterion from [28]:

$$\begin{aligned} \text{MML}_g(\mathbf{y}; \gamma) = & \frac{n-k+2}{2} \left(\ln \frac{n \hat{\tau}_\gamma}{n-k+2} + 1 \right) \\ & + \frac{k-2}{2} \ln \frac{\|\mathbf{X}_\gamma \hat{\beta}_\gamma\|^2}{\max\{k-2, 1\}} + \frac{1}{2} \ln[(n-k)k^2]. \end{aligned}$$

The formula above is applied whenever $\|\mathbf{X}_\gamma \hat{\beta}_\gamma\|^2 / \max\{k-2, 1\} > n \hat{\tau}_\gamma / (n-k+2)$ and $k > 0$. Otherwise, it is used as follows:

$$\text{MML}_g(\mathbf{y}; \emptyset) = \frac{n}{2} \left(\ln \frac{n \hat{\tau}_\gamma}{n-k+2} + 1 \right) + \frac{1}{2} \ln(n-1) + \frac{1}{2}.$$

For completeness, we also consider a second criterion from [28]:

$$\begin{aligned} \text{MML}_u(\mathbf{y}; \gamma) = & \frac{n-k}{2} \ln(2\pi) + \frac{n-k}{2} \left(\ln \frac{n \hat{\tau}_\gamma}{n-k} + 1 \right) + \frac{k}{2} \ln(\pi \mathbf{y}^\top \mathbf{y}) \\ & - \ln \Gamma\left(\frac{k}{2} + 1\right) + \frac{1}{2} \ln(k+1). \end{aligned}$$

Remark that the expression above is for both $k=0$ and $k > 0$.

Next we conduct experiments for simulated and real life data sets.

4.2. Numerical examples

Example 1 illustrates the case of two nested models, which is akin to the model selection problem discussed in Section 3.3. We generate randomly k vectors $\mathbf{z}_1, \dots, \mathbf{z}_k \in \mathbb{R}^{n \times 1}$ such that $\mathbf{z}_i^\top \mathbf{z}_j = \delta_{ij}$ for all $i, j \in \{1, \dots, k\}$, with the convention that $\delta_{\cdot, \cdot}$ denotes the Kronecker operator. In our settings, $k=6$ and $n=50$. Then we choose $\alpha \in (0, \pi/2]$, and define the matrices $\mathbf{X}_{k-1} = [\mathbf{z}_1 \cdots \mathbf{z}_{k-1}]$ and $\mathbf{X}_k = [\mathbf{X}_{k-1} \ \mathbf{z}_k]$.

where $\mathbf{x}_k = \mathbf{z}_1 \cos(\alpha) + \mathbf{z}_k \sin(\alpha)$. It is evident that

$$\mathbf{X}_{k-1}^\top \mathbf{X}_{k-1} = \mathbf{I}, \quad (60)$$

$$\mathbf{X}_k^\top \mathbf{X}_k = \begin{bmatrix} 1 & & \cos\alpha \\ & \ddots & \\ \cos\alpha & & 1 \end{bmatrix}. \quad (61)$$

More importantly, α is the principal angle between the subspaces $\langle \mathbf{X}_{k-1} \rangle$ and $\langle \mathbf{x}_k \rangle$. Given α , we aim to test the performance of various criteria in deciding if the observations are outcomes from $\mathcal{N}(\mathbf{X}_{k-1}\boldsymbol{\beta}_{k-1}, \tau\mathbf{I})$ or from $\mathcal{N}(\mathbf{X}_k\boldsymbol{\beta}_k, \tau\mathbf{I})$. Therefore, we simulate the measurements as follows:

- In the first scenario, we take $\mathbf{y} = \mathbf{X}_{k-1}\boldsymbol{\beta}_{k-1} + \sqrt{\tau}\mathbf{d}$, where $\boldsymbol{\beta}_{k-1} = [1 \dots 1]^\top$, $\tau = (k-1)/n$ and $\mathbf{d} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
- In the second scenario, we have $\mathbf{y} = \mathbf{X}_k\boldsymbol{\beta}_k + \sqrt{\tau}\mathbf{d}$, where $\boldsymbol{\beta}_k = [1 \dots 1]^\top$, $\tau = (k+2\cos\alpha)/n$ and \mathbf{d} has the same significance as above.

Based on (60), the signal-to-noise ratio in the first case is given by $\text{SNR} = \|\mathbf{X}_{k-1}\boldsymbol{\beta}_{k-1}\|^2/(n\tau) = \|\boldsymbol{\beta}_{k-1}\|^2/(n\tau) = 1$. Similarly, by using (61) for the second case, we get

$$\text{SNR} = \frac{\|\mathbf{X}_k\boldsymbol{\beta}_k\|^2}{n\tau} = \frac{k+2\cos\alpha}{n\tau} = 1. \quad (62)$$

For each $\alpha \in \{\pi/180, 2\pi/180, \dots, \pi/2\}$, we generate randomly 500 different realizations of the matrix \mathbf{X}_k by applying the procedure described above. The first $k-1$ columns of each \mathbf{X}_k -matrix define the corresponding \mathbf{X}_{k-1} -matrix. Furthermore, every \mathbf{X}_{k-1} -matrix is used to yield 500 \mathbf{y} -vectors, according to the first scenario. Hence, for each angle α , we have 25×10^4 data vectors which are outcomes from $\mathcal{N}(\mathbf{X}_{k-1}\boldsymbol{\beta}_{k-1}, \tau\mathbf{I})$. Then we decide for each \mathbf{y} if the best model structure is γ_{k-1} or γ_k by employing the eight criteria whose performance is evaluated. In Fig. 1 is plotted the empirical probability of correct estimation versus the angle α . A similar experiment is done for 25×10^4 data vectors simulated, for each α , according to the second scenario. The estimation results are shown in Fig. 2.

In Figs. 1 and 2, we also plot the normalized condition number for the matrix \mathbf{X}_k : $\text{ncond}(\alpha) = \text{cond}(\alpha)/\text{cond}(\alpha_0)$, where $\alpha_0 = \pi/180$. For an arbitrary α , $\text{cond}(\alpha)$ denotes the 2-norm condition number of \mathbf{X}_k , and it equals $[\lambda_{\max}(\alpha)/\lambda_{\min}(\alpha)]^{1/2}$, where $\lambda_{\max}(\alpha)$ and $\lambda_{\min}(\alpha)$ are the maximum and the minimum eigenvalues of the matrix $\mathbf{X}_k^\top \mathbf{X}_k$ [30, p. 78]. It is clear that, for α close to zero, \mathbf{X}_k is badly conditioned numerically. For instance, $\text{cond}(\alpha_0) \approx 115$. However, $\text{cond}(\alpha)$ becomes rapidly smaller when α increases, and we mark in Figs. 1 and 2 the point that corresponds to the value 10 of the 2-norm condition number.

Observe in Fig. 1 that, for all α , SC_{ρ_1} selects the true model with high probability. The fact that the performance of SC_{ρ_1} is not affected by the geometry of the linear subspaces $\langle \mathbf{X}_{k-1} \rangle$ and $\langle \mathbf{x}_k \rangle$ is in line with the results from Section 3.3 (see, for example, Eq. (45) in Lemma 3.1). We remark also in Fig. 1 that the behavior of MML_g , MML_u , BIC and AIC_{c3} is very similar with that of SC_{ρ_1} .

The relationship between the performance of SC_{ρ_2} and SC_{ρ_1} can be understood better by recalling that, according to Proposition 3.4, the difference of the expectations of penalty terms, $\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_2)] - \mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)]$, is positive only

for $\alpha \in (0, \alpha^*)$, it decreases as long as $\alpha \leq \alpha_{\min}$, and then increases when $\alpha \in (\alpha_{\min}, \pi/2)$. This is very well reflected by the graphs within Fig. 1, where SC_{ρ_2} is slightly better than SC_{ρ_1} when $\alpha = \pi/180$, but its performance declines when α increases and, after reaching a minimum point, SC_{ρ_2} improves such that it becomes as good as SC_{ρ_1} when $\alpha = \pi/2$.

From the identities (49) and (62), we get $\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)] = (k+2\cos\alpha)(1+k/n)$. So, we expect that, with our experimental settings, the criterion SC_{ρ_1} penalizes less the γ_k -model when α increases. This theoretical result, which is based on Lemma 3.2, agrees perfectly with the empirical results shown in Fig. 2.

By looking simultaneously at Figs. 1 and 2, we note that SC_{ρ_2} prefers the γ_k -model when the condition number takes large values, and this effect is undesirable. On contrary, SC_{ρ_3} selects the γ_{k-1} -model whenever the condition number is high, which shows that SC_{ρ_3} is prone to choose the model whose explanatory variables are linearly independent, and not the “true” model. When $\alpha = \pi/2$, or equivalently the matrix \mathbf{X}_k is orthonormal, the criteria SC_{ρ_1} , SC_{ρ_2} and SC_{ρ_3} reduce to one single criterion, as we know from Proposition 3.1.

In Fig. 1, CME has the poorest results as it strongly prefers the γ_k -model. This can be explained by noticing in (59) that $\frac{1}{2}\ln|\mathbf{X}_{k-1}^\top \mathbf{X}_{k-1}|$ is a penalty term for the γ_{k-1} -model, and $\frac{1}{2}\ln|\mathbf{X}_k^\top \mathbf{X}_k|$ is a penalty term for the γ_k -model. In our settings, $\frac{1}{2}\ln|\mathbf{X}_{k-1}^\top \mathbf{X}_{k-1}| = 0$, whereas $\frac{1}{2}\ln|\mathbf{X}_k^\top \mathbf{X}_k| \rightarrow -\infty$ when $\alpha \rightarrow 0$. It is worth mentioning that $\ln|\mathbf{X}_k^\top \mathbf{X}_k|$ is also a penalty term within SC_{ρ_2} -formula in (31). However, the significant decrease of $\ln|\mathbf{X}_k^\top \mathbf{X}_k|$ when $\alpha \rightarrow 0$ is compensated in SC_{ρ_2} -formula by the increase of the term $k\ln(\|\boldsymbol{\beta}_k\|^2/n/k)$. More interestingly, CME has difficulties in correctly identifying the γ_{k-1} -model even when α takes values close to $\pi/2$. The reason is that the logarithm of determinant of the observed FIM is not guaranteed to be a correct penalty term even if the columns of \mathbf{X} are almost orthogonal. We will investigate more carefully this aspect in the next example.

Example 2 is taken from [16] and is focused on the variable selection for the linear regression in (1), when the matrix \mathbf{X} has the particular form

$$\mathbf{X} = \begin{bmatrix} \cos(2\pi f_1) & \cdots & \cos(2\pi f_8) \\ \vdots & \ddots & \vdots \\ \cos[2\pi f_1(N-1)] & \cdots & \cos[2\pi f_8(N-1)] \end{bmatrix},$$

where $f_j = [0.10 + (j-1)/100]$ for $j \in \{1, \dots, 8\}$. With the notations from (1), $n = N-1$ and $m=8$. The vector of linear parameters $\boldsymbol{\beta}$ contains the unknown amplitudes, and the variance of the additive Gaussian noise is assumed to be unknown. The competitors are eight nested models with structures $\gamma_1, \dots, \gamma_8$, where $\gamma_k = \{1, \dots, k\}$. Equivalently, the regressor matrix \mathbf{X}_{γ_k} for the model γ_k , $k \in \{1, \dots, 8\}$, is given by the first k columns of \mathbf{X} . For simplicity, we use the notation \mathbf{X}_k instead of \mathbf{X}_{γ_k} , and $\boldsymbol{\beta}_k$ instead of $\boldsymbol{\beta}_{\gamma_k}$.

To mimic the experiments from [16], we simulate data according to the structure γ_3 by taking $\boldsymbol{\beta}_3 = [1 \ 1 \ 1]^\top$. In the first experiment, the noise variance is $\tau = 10$ and the sample size is varied by choosing N from the set $\{100, 110, \dots, 300\}$. In the second experiment, the sample size is kept fixed

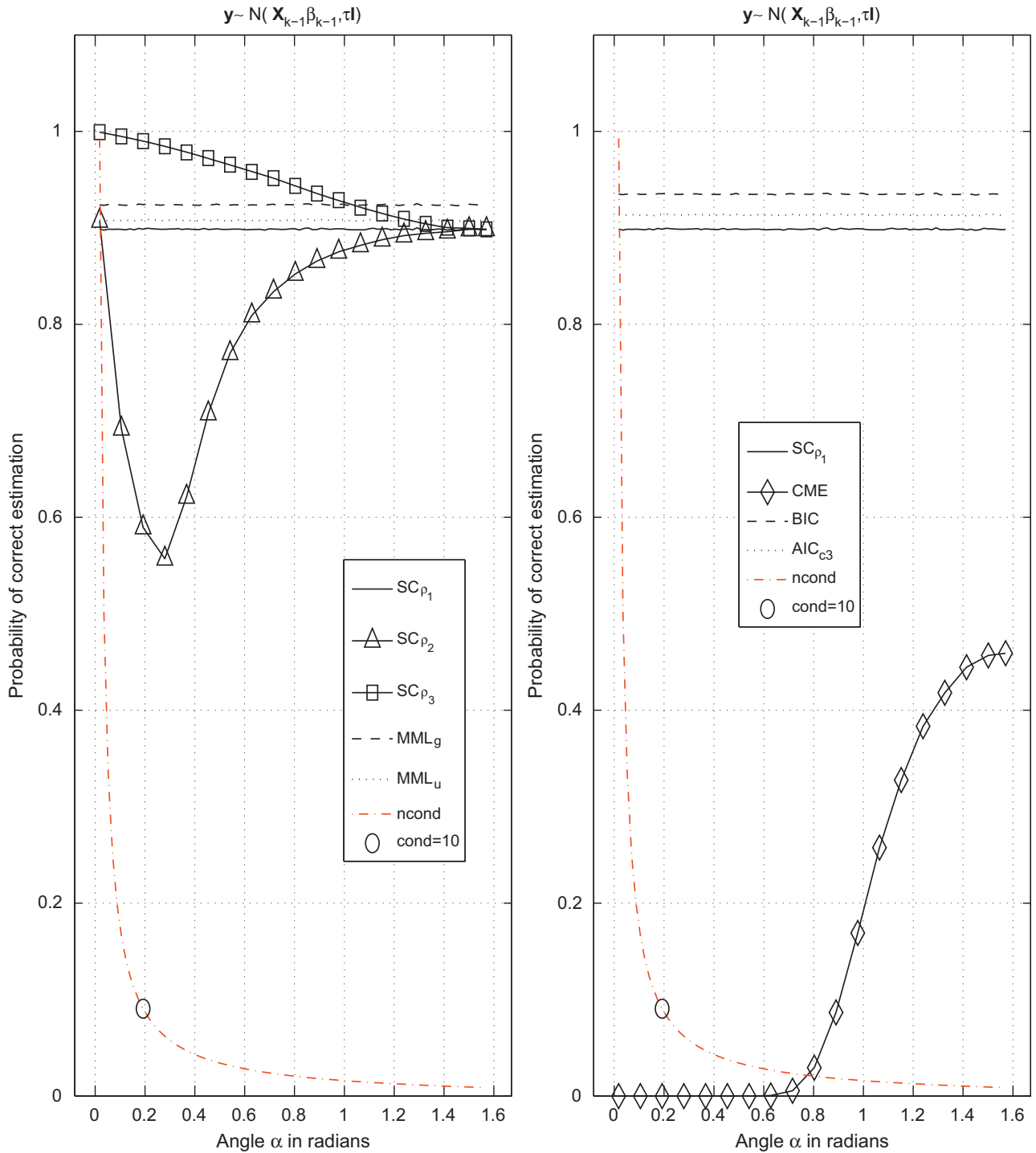


Fig. 1. Example 1—the empirical probability of deciding correctly that the observations $\mathbf{y} \in \mathbb{R}^n$ are outcomes from $\mathcal{N}(\mathbf{X}_{k-1}\boldsymbol{\beta}_{k-1}, \tau\mathbf{I})$, and not from $\mathcal{N}(\mathbf{X}_k\boldsymbol{\beta}_k, \tau\mathbf{I})$. With the convention that $\mathbf{X}_k = [\mathbf{X}_{k-1} \ \mathbf{x}_k]$, α denotes the principal angle between the linear subspaces $\langle \mathbf{X}_{k-1} \rangle$ and $\langle \mathbf{x}_k \rangle$. For an arbitrary α , $\text{cond}(\alpha)$ denotes the 2-norm condition number of \mathbf{X}_k . The normalized condition number is $\text{ncond}(\alpha) = \text{cond}(\alpha)/\text{cond}(\alpha_0)$, where $\alpha_0 = \pi/180$. For the simulated data, $n = 50$, $k = 6$, and τ is chosen such that $\text{SNR} = 0$ dB.

($N=100$), and the SNR is varied by modifying the noise variance such that $1/\tau \in \{0.01, 0.02, \dots, 0.2\}$. The empirical probabilities of selecting correctly the number of sinusoids are plotted in Fig. 3 for the first experiment, and in Fig. 4 for the second experiment. Note that the probabilities shown in Fig. 3 are obtained, for each value of N , from 10^4 runs.

Similarly, in the second experiment, the number of runs for each value of $1/\tau$ is 10^4 .

In both figures, the graphs for SC_{p_1} , SC_{p_2} and SC_{p_3} almost coincide. This is because [5,16,14]

$$\mathbf{X}_k^T \mathbf{X}_k \approx (n/2)\mathbf{I}, \quad \forall k \in \{2, \dots, 8\}, \quad (63)$$

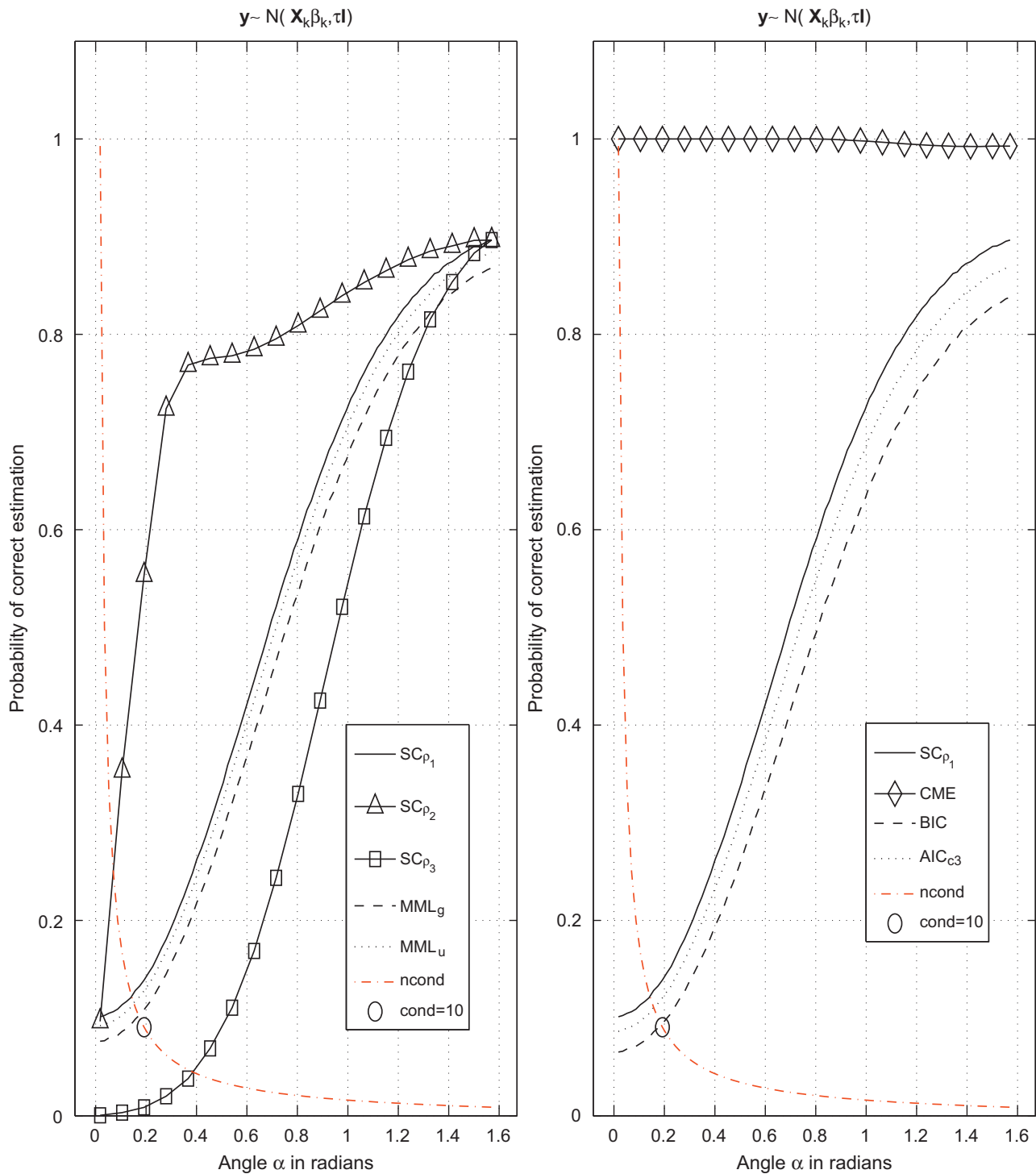


Fig. 2. Example 1—the empirical probability of deciding correctly that the observations $\mathbf{y} \in \mathbb{R}^n$ are outcomes from $\mathcal{N}(\mathbf{X}_k \beta_k, \tau \mathbf{I})$, and not from $\mathcal{N}(\mathbf{X}_{k-1} \beta_{k-1}, \tau \mathbf{I})$. All conventions are the same like in Fig. 1.

which makes the condition within point (a) of Proposition 3.1 to be satisfied. Additionally, MML_g and MML_u perform similarly with SC_{p_1} , and they are both superior to SC_{p_1} only when $N > 200$ as we can see in Fig. 3. We observe in the same figure that AIC_{c3} outperforms other criteria when $N < 150$. The good estimation capabilities of AIC_{c3} when sample size is small can be noticed also in Fig. 4 where, for

$N = 100$, AIC_{c3} is superior to SC_{p_1} and BIC for almost all SNRs. On contrary, when $N > 200$, the estimation results of AIC_{c3} are modest, and BIC improves significantly. The reason is simple: AIC_{c3} has been designed especially for the small sample case [32], whereas the use of BIC is recommended for large samples because its derivation relies on asymptotic approximations [29]. It is remarkable that SC_{p_1} is nearly as

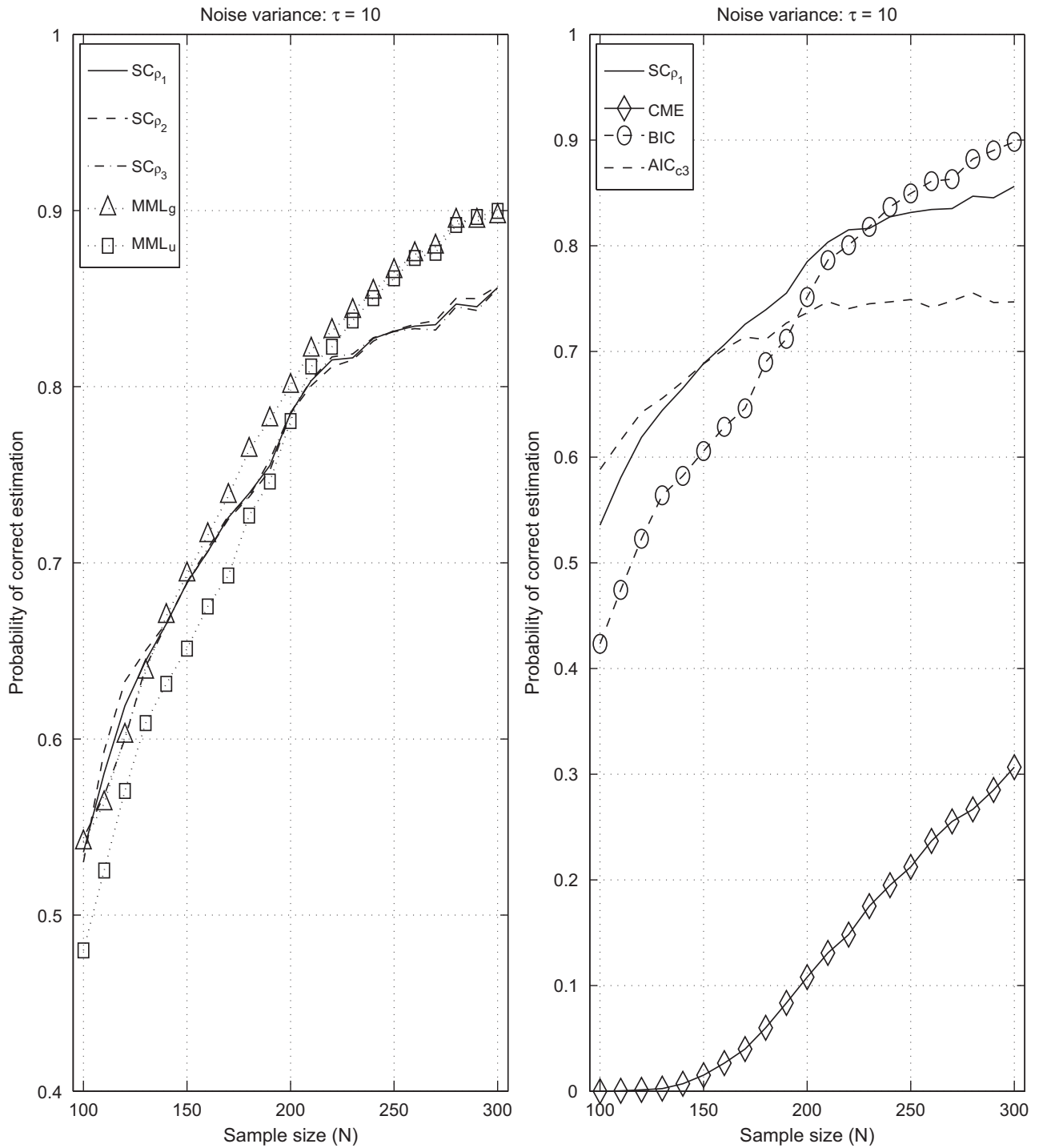


Fig. 3. Example 2—the empirical probability of estimating correctly the number of sinusoids versus the sample size. Note that the range of values being presented along the vertical axes is different for the two plots.

good as AIC_{c3} when N is small, and it is only marginally inferior to BIC when N is large.

The performance of CME is again very modest, and it can be explained by re-writing, in a more convenient form, the expression from (59). We approximate $\ln \Gamma((n-k)/2)$ by (26), and then we neglect the sum $(n/2) \ln[2\pi n \exp(1)] - \frac{1}{2} \ln(4\pi n^2)$ which does not depend

on k . So, we obtain the following formula when the structure is γ_k :

$$CME(\mathbf{y}; \gamma_k) = \frac{n}{2} \ln \hat{\tau}_k + \frac{1}{2} \ln \left| \frac{(n-k)/n}{2\pi \exp(1) \hat{\tau}_k} \mathbf{X}_k^T \mathbf{X}_k \right| - \left[\ln \hat{\tau}_k + \frac{n-3}{2} \ln(n-k) \right]. \quad (64)$$

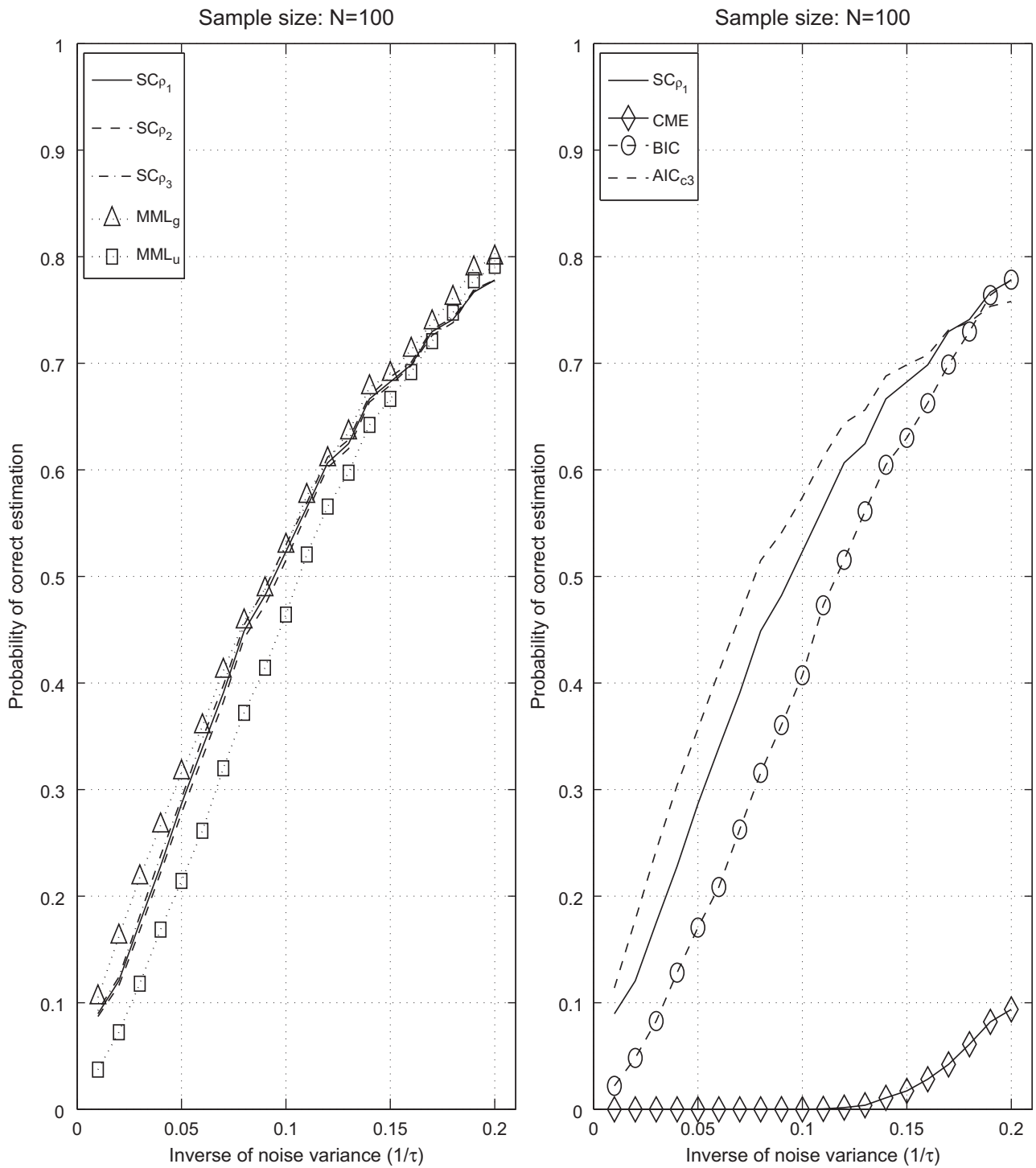


Fig. 4. Example 2—the empirical probability of estimating correctly the number of sinusoids versus the inverse of the noise variance.

It is obvious that $\hat{\tau}_k = \|(\mathbf{I} - \mathbf{P}_k)\mathbf{y}\|^2/n$, where \mathbf{P}_k is the orthogonal projection matrix onto the linear subspace $\langle \mathbf{X}_k \rangle$. By using (63), we notice that the second term within (64) is given by $\text{PEN}(\mathbf{y}; \gamma_k) = (k/2)\ln[(n-k)/(4\pi \exp(1)\hat{\tau}_k)]$. For small n , $\text{PEN}(\mathbf{y}; \gamma_k)$ does not increase fast enough when the model order k becomes larger. For comparison, note in

(56) that the BIC penalty term is $(k/2)\ln n$. The fact that the penalty of CME is possibly incorrect for small sample size has been already analyzed in the case when the variance of the Gaussian noise is a priori known (see [16]). However, we show in the next example that CME is rather good in estimating the order of a polynomial model.

Example 3 is also taken from [16], and this time the regressor matrix is given by

$$\mathbf{X} = \begin{bmatrix} 1^0 & 1^1 & \dots & 1^9 \\ 2^0 & 2^1 & \dots & 2^9 \\ \vdots & \vdots & \ddots & \vdots \\ (N-1)^0 & (N-1)^1 & \dots & (N-1)^9 \end{bmatrix}.$$

It is evident that $n = N - 1$ and $m = 10$. Similarly with the previous example, the number of competing nested models equals m and their structures are such that $\gamma_k = \{1, \dots, k\}$ for all $k \in \{1, \dots, m\}$. The variance of the additive Gaussian noise is assumed to be unknown, and we use again the notation \mathbf{X}_k instead of \mathbf{X}_{γ_k} , and β_k instead of β_{γ_k} .

The data are simulated according to the structure γ_3 such that the linear parameters are $\beta_3 = [0 \ 0.4 \ 0.1]^T$. Hence, the

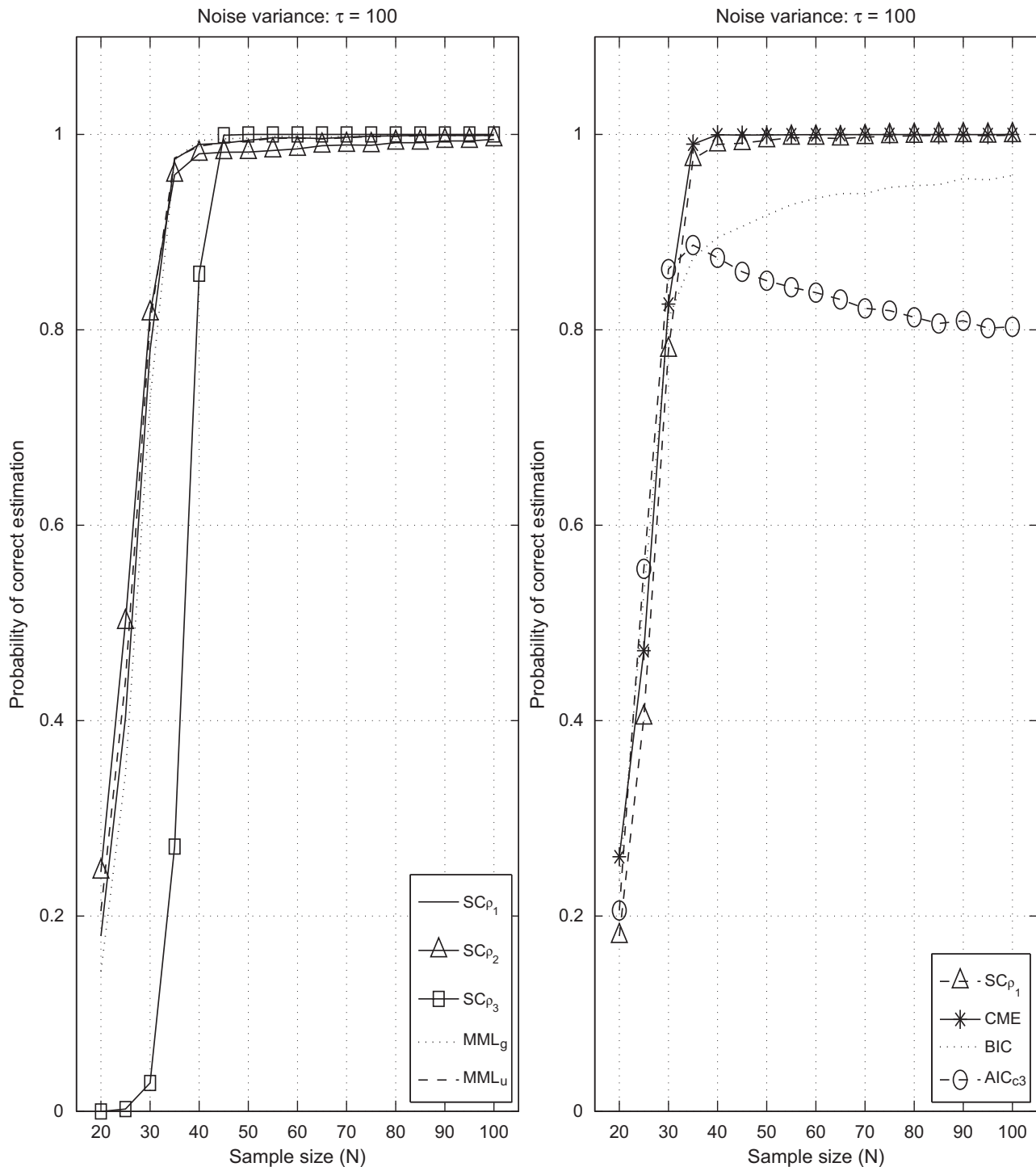


Fig. 5. Example 3—the empirical probability of estimating correctly the order of the polynomial model versus the sample size.

observations represent a parabolic signal in noise. In the first scenario, the noise variance is $\tau = 100$ and the sample size is varied by choosing N from the set $\{20, 25, \dots, 100\}$. Based on 10^4 trials for each value of N , we evaluate the empirical probabilities of selecting the γ_3 -structure, and we plot them in Fig. 5. Then the sample size is kept fixed ($N=40$), and the SNR is varied by modifying the noise

variance such that $1/\tau \in \{1/10^3, 2/10^3, \dots, 10/10^3\}$. The number of runs for each value of $1/\tau$ is 10^4 , and the results are shown in Fig. 6.

Remark in Fig. 5 that the results of SC_{ρ_1} , SC_{ρ_2} , MML_g and MML_u are very similar for all values of N , whereas SC_{ρ_3} fails to estimate properly the structure when $N \leq 40$. Moreover, for $N=40$, SC_{ρ_3} is inferior to SC_{ρ_1} , SC_{ρ_2} , MML_g

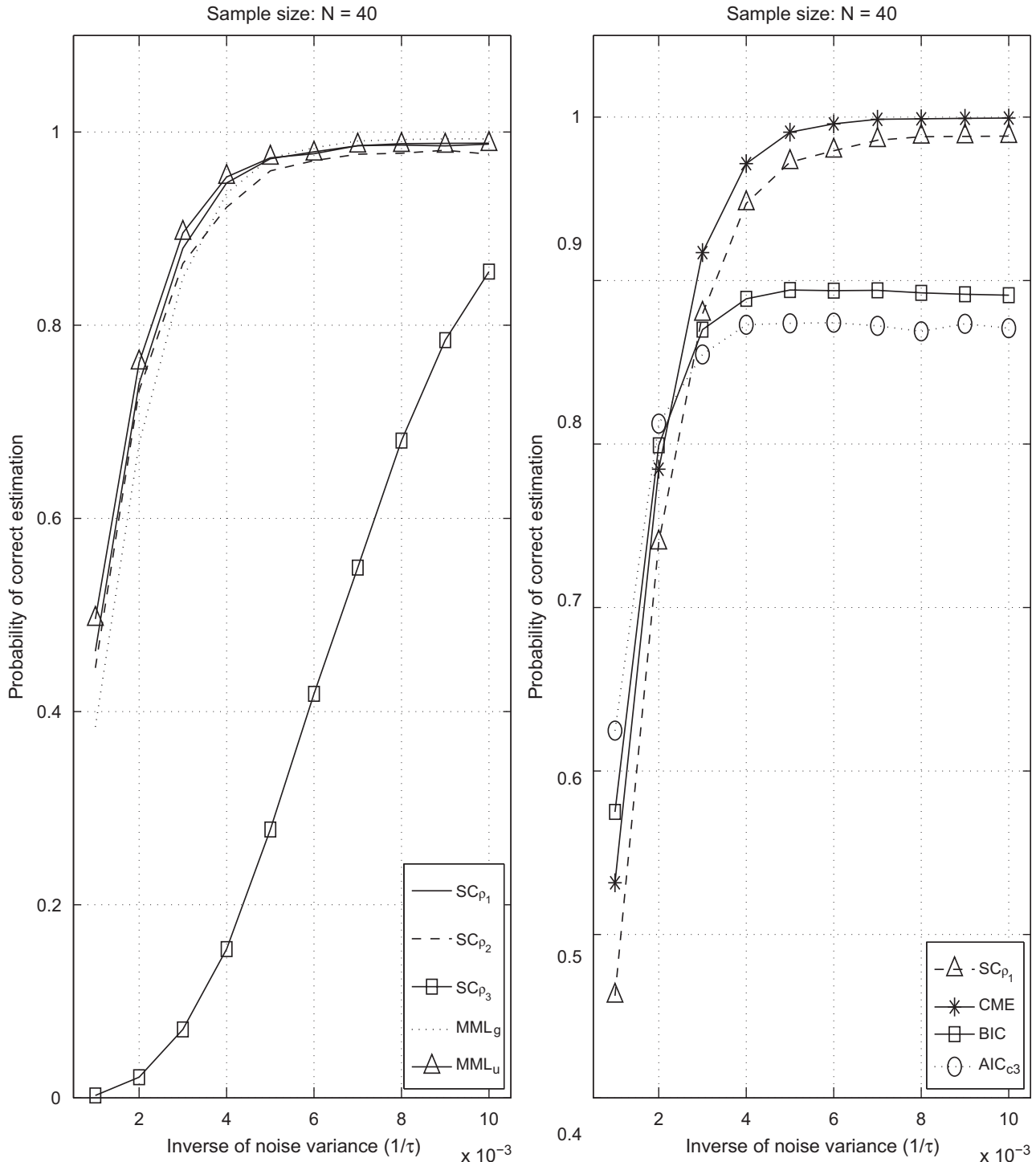


Fig. 6. Example 3—the empirical probability of estimating correctly the order of the polynomial model versus the inverse of the noise variance. Note that the range of values being presented along the vertical axes is different for the two plots.

and MML_u for all values of SNR considered in Fig. 6. CME performs extremely well in both figures, and SC_{ρ_1} is almost as good as CME. AIC_{c3} confirms in Fig. 5 what we have already seen in the previous example: it outperforms other criteria when the sample size is small ($N \leq 30$), but for large sample size its estimation capabilities are modest. The accuracy of the BIC estimate is better and better when N increases, but even for $N=100$, BIC remains inferior to CME. In Fig. 6, CME outperforms BIC for a large span of SNR values.

The fact that, for the polynomial model, CME is superior to BIC has been already pointed out in [15,16], and it can be understood by resorting to the following asymptotic results from [5]:

$$\mathbf{X}_k^\top \mathbf{X}_k \approx \begin{bmatrix} N & \frac{N^2}{2} & \cdots & \frac{N^k}{k} \\ \frac{N^2}{2} & \frac{N^3}{3} & \cdots & \frac{N^{k+1}}{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{N^k}{k} & \frac{N^{k+1}}{k+1} & \cdots & \frac{N^{2k-1}}{2k-1} \end{bmatrix}, \quad (65)$$

$$|\mathbf{X}_k^\top \mathbf{X}_k| = O(N^{k^2}). \quad (66)$$

Therefore, $\frac{1}{2} \ln |\mathbf{X}_k^\top \mathbf{X}_k|$ which is the penalty term of CME can be written as $[(k^2/2) \ln n + O(1)]$. This shows immediately that $(k/2) \ln n$, the penalty term of BIC, is not the correct one (see [5] for a more detailed discussion). More interestingly, by combining (34) with the approximations from (65)–(66), we have

$$D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1) = \hat{\beta}_k^\top (\mathbf{X}_k^\top \mathbf{X}_k) \hat{\beta}_k = O(N^{2k-1}),$$

$$D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_2) = \frac{\|\hat{\beta}_k\|^2}{|\mathbf{X}_k^\top \mathbf{X}_k|^{-1/k}} = O(N^k),$$

$$D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_3) = \frac{\hat{\beta}_k^\top (\mathbf{X}_k^\top \mathbf{X}_k)^2 \hat{\beta}_k}{|\mathbf{X}_k^\top \mathbf{X}_k|^{1/k}} = O(N^{3k-2}).$$

According to (33), the penalty term of $\text{SC}_{\rho_i}(\mathbf{y}; \gamma_k)$ is given by $\ln D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_i)$, where $i \in \{1, 2, 3\}$. Thus, we can express as follows the penalty terms of the criteria listed below:

$$\text{SC}_{\rho_1} : (2k^2 - k) \ln n + O(1),$$

$$\text{SC}_{\rho_2} : k^2 \ln n + O(1),$$

$$\text{SC}_{\rho_3} : (3k^2 - 2k) \ln n + O(1).$$

Recall that the formula in (33) was multiplied by two for writing the equations in a more compact form. Consequently, the above results must be divided by two before comparing them with the penalty term of CME. Note that only SC_{ρ_2} penalizes the complexity of the model as CME does. SC_{ρ_3} is the criterion that deviates the most from the recommended penalty which is $[(k^2/2) \ln n + O(1)]$, and this explains why, in Figs. 5 and 6, the performance of SC_{ρ_3} is modest.

The experimental results obtained for Examples 1–3 lead to some guidance on the application of various model selection criteria to the estimation problems which have been investigated. We summarize the recommendations in Table 1.

Example 4 is focused on the predictive capabilities of the model selection criteria which are investigated. Given a data set that contains, for n different instances, the measurements of m input attributes along with the measurements of the response variable, we randomly choose n_{tr} samples to be the training set. Based on the linear regression model, each criterion uses the training set to choose the most relevant input attributes. The model learned by each criterion is applied to the remaining $n - n_{tr}$ samples, which constitute the test set, and the squared prediction error is computed.

The data sets used in our experiments are listed below. For each of them, we indicate the values of n and m , as well as the repository where they are publicly available:

1. *Housing* data set: $n=506$, $m=13$, <http://archive.ics.uci.edu/ml/datasets/Housing>.

Table 1

Guidance on the use of the eight criteria for the estimation problems in Examples 1–3: **A**—recommended; **B**—acceptable; **C**—unsatisfactory; **D**—not recommended. For each example, the information about the experimental conditions (sample size and SNR) is provided with the conventions from Figs. 1 to 6.

Experimental conditions	SC_{ρ_1}	SC_{ρ_2}	SC_{ρ_3}	MML_g	MML_u	CME	BIC	AIC_{c3}
Example 1 —Select the variables which are linearly independent								
$n=50$; SNR=0 dB	B	C	A	B	B	D	B	B
Example 2 —Estimate correctly the number of sinusoids embedded in Gaussian noise								
$N \in [100, 150]$; $\tau^{-1} = 0.10$	B	B	B	B	C	D	C	B
$N \in [150, 200]$; $\tau^{-1} = 0.10$	B	B	B	B	B	D	B	B
$N \in (200, 300]$; $\tau^{-1} = 0.10$	B	B	B	A	A	D	A	C
$N=100$; $\tau^{-1} \in [0.01, 0.07]$	C	C	C	C	C	D	C	C
$N=100$; $\tau^{-1} \in (0.07, 0.10)$	C	C	C	C	C	D	C	B
$N=100$; $\tau^{-1} \in [0.10, 0.12]$	B	B	B	B	C	D	C	B
$N=100$; $\tau^{-1} \in (0.12, 0.20]$	B	B	B	B	B	D	B	B
Example 3 —Estimate correctly the order of a polynomial in Gaussian noise								
$N \in [20, 25]$; $\tau^{-1} = 0.01$	C	C	D	C	C	C	C	C
$N \in (25, 40]$; $\tau^{-1} = 0.01$	B	B	D	B	B	B	B	B
$N=40$; $\tau^{-1} = 0.01$	A	A	C	A	A	A	C	C
$N \in (40, 100]$; $\tau^{-1} = 0.01$	A	A	A	A	A	A	C	D
$N=40$; $\tau^{-1} \in [0.001, 0.004]$	B	B	D	B	B	B	B	B
$N=40$; $\tau^{-1} \in [0.004, 0.01)$	A	A	D	A	A	A	C	C

2. *Diabetes* data set (standardized): $n=442$, $m=10$, <http://www-stat.stanford.edu/~hastie/Papers/LARS>.
3. *Concrete* compressive strength data set: $n=1030$, $m=8$, <http://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength>.

At the addresses outlined above, the interested reader can find the data tables, an accurate description of their content, along with references to previous works where they have been utilized. For instance, all three data sets have been also used for the experimental part of [28]. We note that, in [28], the Housing data set has been altered as follows: the measurements corresponding to the attribute CHAS (Charles River dummy variable) have been removed, the values of the attribute NOX (nitric oxides concentration—parts per 10 million) have been multiplied by 100, and the \mathbf{y} -vector for the response variable has been transformed such that to have zero-mean. Similarly, the vector for the response variable within Concrete compressive strength data set has been modified to have zero-mean. Because we want our settings to be like in [28], we apply the same changes. Moreover, the model selection during the training step is slightly different than how it was performed in Examples 1–3:

- One modification is that we select the best model among all γ -structures which are subsets of $\{1, \dots, m\}$, including the case $\gamma = \emptyset$. Therefore, the tested models are not nested, and we cannot any longer apply the recursive least-squares algorithm [14, p. 237] to estimate the linear parameters, as it was done in Examples 1–3. Like in [28], we use the Moore–Penrose pseudoinverse.
- Another modification is that we do *not* neglect the term $L(\gamma)$ which quantifies the complexity of the structure. To be in line with [28], we do not apply the formula from [26], but the following one:

$$L'(\gamma) = \ln \binom{m}{k} + \ln(m+1),$$

where k denotes the cardinality of γ . Obviously, $2L'(\gamma)$ is added to SC_{ρ_1} , SC_{ρ_2} , SC_{ρ_3} , and $L'(\gamma)$ is added to the other criteria.

The predictive accuracy is evaluated for five different values of n_{tr} , and the results are shown in Table 2. Note that each entry within Table 2 is calculated as an average of the prediction errors obtained from 10^3 random partitions of the data sets into training/test subsets. The results for MML_g and MML_u are identical with those from [28]. Because in [28], it was not used the Stirling approximation (26) when evaluating SC_{ρ_1} , for this criterion, there exist small differences between the results from Table 2 and the results reported in [28].

Based on the empirical evidence, it is not possible to decide that one particular criterion has stronger prediction capabilities than the others. It is interesting to remark in Table 2 that it does not exist any combination of experimental settings for which BIC yields the smallest prediction error. The same is true for SC_{ρ_1} . Overall, SC_{ρ_2} is slightly superior to SC_{ρ_1} . CME performs surprisingly well for the Diabetes data set, but for the other two data sets, its results are moderate.

5. Conclusions

In the case of the Gaussian linear regression, the parametric complexity is not finite and the only possibility for obtaining NML-based selection rules is to constrain the data space. Even if this was recognized long time ago, the solutions proposed so far are only punctual results which treat some particular constraints. In this paper, we have introduced a general methodology for addressing the problem. Based on the new findings, we demonstrated how the rhomboidal constraint yields a new NML-based formula. Additionally, we used the ellipsoidal constraint to re-derive three criteria that have been introduced in the previous literature: SC_{ρ_1} [25] and SC_{ρ_2} and SC_{ρ_3} [18]. They have been compared against BIC [29], AIC_{c3} [32], CME [15] and MML_g and MML_u [28].

The theoretical analysis and the Monte Carlo simulations led to the following outcomes: (a) SC_{ρ_3} has the strongest tendency to select the variables which are linearly independent; (b) SC_{ρ_1} , SC_{ρ_2} and SC_{ρ_3} reduce to one single criterion when the regression matrix is orthonormal; (c) MML_g and

Table 2

Example 4—squared prediction errors obtained for real life measurements. For all data sets, it is written in bold the best result for each n_{tr} .

Data set	n_{tr}	SC_{ρ_1}	SC_{ρ_2}	SC_{ρ_3}	MML_g	MML_u	CME	BIC	AIC_{c3}
Housing	25	69.976	52.529	53.249	61.922	71.509	85.282	70.326	59.463
	50	36.933	35.265	37.268	36.340	36.635	36.147	36.511	36.577
	100	29.323	30.210	30.523	29.624	29.383	29.079	29.516	28.343
	200	26.023	27.711	27.657	26.424	26.162	26.897	26.535	25.271
	400	24.315	25.998	26.225	24.304	24.299	24.645	24.365	24.321
Diabetes	25	4824.3	4362.9	4553.0	4445.0	4819.2	5386.5	4647.5	4506.3
	50	3855.3	3645.3	3902.0	3851.2	3843.8	3722.5	3819.5	3743.1
	100	3355.2	3259.9	3410.1	3385.3	3364.2	3237.2	3368.4	3301.5
	200	3165.9	3099.7	3210.7	3199.6	3173.3	3069.5	3195.4	3073.4
	400	3046.9	3060.5	3053.4	3052.8	3052.7	3026.9	3055.7	3026.9
Concrete	25	225.18	257.71	245.86	221.2	227.41	279.27	235.07	245.40
	50	148.67	148.57	147.11	147.46	149.25	162.36	150.06	148.78
	100	123.82	121.56	121.59	122.90	123.65	124.00	123.29	124.11
	200	114.56	113.89	114.05	114.37	114.50	114.17	114.31	114.89
	400	111.67	111.12	111.46	111.59	111.64	111.22	111.56	111.70

MML_u perform similarly with SC_{ρ₁} and they are superior to SC_{ρ₁} for some particular experimental settings; (d) AIC_{c3} is very good when the sample size is small, but it has modest results when the sample size is large; (e) BIC has a behavior which is opposite to that of AIC_{c3}, and the performance of SC_{ρ₁} is an excellent compromise between BIC and AIC_{c3}; (f) CME poses troubles for some models, but in the case of the polynomial model, it is ranked the first for a large range of sample sizes; (g) SC_{ρ₁}, SC_{ρ₂}, MML_g and MML_u are nearly as good as CME for the polynomial model, while SC_{ρ₃} has difficulties in this case.

Acknowledgments

The work of C.D. Giurcăneanu and S.A. Razavi was supported by the Academy of Finland, Project Nos. 113572, 118355, 134767 and 213462. The work of A. Liski was supported by the Academy of Finland under Project No. 213462. S.A. Razavi is grateful to Dr. Enes Makalic for providing the Matlab code used in the experimental part of [28].

Appendix A. Evaluation of the normalized maximum likelihood

The techniques that we apply in this section are very similar with those used in [17,18,25–27].

Computation of $C_\rho(R, \tau_0)$. First we note that the estimated parameter vector $[\hat{\beta}^\top \hat{\tau}]^\top$ is a sufficient statistic, and the density $f(\mathbf{y}; \beta, \tau)$ can be factored as follows:

$$f(\mathbf{y}; \beta, \tau) = f(\mathbf{y}|\hat{\beta}, \hat{\tau})g(\beta, \tau; \hat{\beta}, \hat{\tau}), \quad (\text{A.1})$$

where $f(\mathbf{y}|\hat{\beta}, \hat{\tau})$ does not depend on the unknowns β and τ . According to [31, Theorem 3.5], the estimates $\hat{\beta}$ and $\hat{\tau}$ are statistically independent, and we have

$$g(\beta, \tau; \hat{\beta}, \hat{\tau}) = g_1(\hat{\beta}; \beta, \tau)g_2(\hat{\tau}; \tau),$$

$$g_1(\hat{\beta}; \beta, \tau) = \frac{|\mathbf{X}^\top \mathbf{X}|^{1/2}}{(2\pi\tau)^{k/2}} \exp\left(-\frac{\|\mathbf{X}(\hat{\beta} - \beta)\|^2}{2\tau}\right),$$

$$g_2(\hat{\tau}; \tau) = \frac{n^{(n-k)/2}}{\Gamma(\frac{n-k}{2})2^{(n-k)/2}} \left(\frac{\hat{\tau}}{\tau}\right)^{(n-k)/2} \frac{1}{\hat{\tau}} \exp\left(-\frac{n\hat{\tau}}{2\tau}\right).$$

By employing (15), we obtain

$$g(\hat{\beta}, \hat{\tau}; \hat{\beta}, \hat{\tau}) = A_{n,k} \hat{\tau}^{-k/2-1}. \quad (\text{A.2})$$

Then we define $\mathcal{P}(R, \tau_0) = \{[\hat{\beta}^\top \hat{\tau}]^\top : \rho(\hat{\beta}) \leq R, \hat{\tau} \geq \tau_0\}$ and $\mathcal{Y}(\hat{\beta}, \hat{\tau}) = \{\mathbf{y} : \hat{\beta}(\mathbf{y}) = \hat{\beta}, \hat{\tau}(\mathbf{y}) = \hat{\tau}\}$. After these preparations, we evaluate the integral in (12):

$$C_\rho(R, \tau_0) = \int_{\mathcal{P}(R, \tau_0)} \left[\int_{\mathcal{Y}(\hat{\beta}, \hat{\tau})} f(\mathbf{y}|\hat{\beta}, \hat{\tau}) d\mathbf{y} \right] g(\hat{\beta}, \hat{\tau}; \hat{\beta}, \hat{\tau}) d\hat{\beta} d\hat{\tau} \quad (\text{A.3})$$

$$= A_{n,k} \int_{\tau_0}^{\infty} \hat{\tau}^{-k/2-1} d\hat{\tau} \int_{\mathcal{B}(R)} d\hat{\beta} = (2A_{n,k}/k) \tau_0^{-k/2} V_\rho(R). \quad (\text{A.4})$$

Remark in (A.3) that the inner integral gives unity [26]. The use of (A.2) and some simple manipulations yield (A.4). Additionally, (10) and (A.4) lead to (14).

Computation of $\bar{C}_\rho(R_1, R_2, \tau_1, \tau_2)$. For evaluating the normalizing constant in (18), we define $\mathcal{P}(R_1, R_2, \tau_1, \tau_2) = \{[\hat{\beta}^\top \hat{\tau}]^\top : R_1 \leq \rho(\hat{\beta}) \leq R_2, \tau_1 \leq \hat{\tau} \leq \tau_2\}$ and $\mathcal{B}(R_1, R_2) = \{\hat{\beta} : R_1 \leq \rho(\hat{\beta}) \leq R_2\}$. So,

$$\bar{C}_\rho(R_1, R_2, \tau_1, \tau_2) = \int_{\mathcal{Y}(R_1, R_2, \tau_1, \tau_2)} \hat{f}(\mathbf{y}; \hat{R}, \hat{\tau}_0) d\mathbf{y} \\ = \int_{\mathcal{Y}(R_1, R_2, \tau_1, \tau_2)} \frac{f(\mathbf{y}|\hat{\beta}, \hat{\tau})g(\hat{\beta}, \hat{\tau}; \hat{\beta}, \hat{\tau})}{C_\rho(\hat{R}, \hat{\tau}_0)} d\mathbf{y} \quad (\text{A.5})$$

$$= \int_{\mathcal{P}(R_1, R_2, \tau_1, \tau_2)} \frac{g(\hat{\beta}, \hat{\tau}; \hat{\beta}, \hat{\tau})}{C_\rho(\hat{R}, \hat{\tau}_0)} \left[\int_{\mathcal{Y}(\hat{\beta}, \hat{\tau})} f(\mathbf{y}|\hat{\beta}, \hat{\tau}) d\mathbf{y} \right] d\hat{\beta} d\hat{\tau} \quad (\text{A.6})$$

$$= \int_{\mathcal{P}(R_1, R_2, \tau_1, \tau_2)} \frac{A_{n,k} \hat{\tau}^{-k/2-1}}{A_{n,k} (2/k) \hat{\tau}^{-k/2} V_\rho(\hat{R})} d\hat{\beta} d\hat{\tau} \quad (\text{A.7})$$

$$= \frac{k}{2} \int_{\tau_1}^{\tau_2} \frac{1}{\hat{\tau}} d\hat{\tau} \int_{\mathcal{B}(R_1, R_2)} \eta^{-1} [\rho(\hat{\beta})]^{-\zeta k} d\hat{\beta} \quad (\text{A.8})$$

$$= \frac{k}{2} \ln \frac{\tau_2}{\tau_1} \int_{R_1}^{R_2} \frac{(\eta \zeta k) R^{\zeta k-1}}{\eta R^{\zeta k}} dR \\ = \frac{\zeta k^2}{2} \ln \frac{\tau_2}{\tau_1} \ln \frac{R_2}{R_1}.$$

Note that in (A.5) we use again the factorization from (A.1). Similarly with (A.3), the inner integral in (A.6) gives unity. The identity in (A.7) is derived straightforwardly from (A.2), (A.4) and (A.6). For the calculation of the second integral in (A.8), we apply the same technique as in [25,26] and, based on (10), we take the element of volume to be $dV_\rho = \eta \zeta k R^{\zeta k-1} dR$. After some simple algebra, we get the result in (19).

Evaluation of the approximate formula (23). Note that the approximation from (23) can be applied for a much more general class of models, and not only for the model in (2). The proof given in [24] treats the general case and is based on sophisticated mathematical derivations. However, it was already pointed out in [26, Section 5.2.2] that the proof can be simplified if the analyzed model satisfies a particular condition. With our notations, the condition is as follows:

$$\lim_{n \rightarrow \infty} \frac{g(\hat{\beta}, \hat{\tau}; \hat{\beta}, \hat{\tau})}{[n/(2\pi)]^{(k+1)/2} |\mathbf{J}_n(\hat{\beta}, \hat{\tau})|^{1/2}} = 1. \quad (\text{A.9})$$

Observe that Eq. (25) leads to

$$\left(\frac{n}{2\pi}\right)^{(k+1)/2} |\mathbf{J}_n(\hat{\beta}, \hat{\tau})|^{1/2} = \check{A}_{n,k} \hat{\tau}^{-k/2-1},$$

$$\check{A}_{n,k} = \frac{|\mathbf{X}^\top \mathbf{X}|^{1/2} \sqrt{n}}{(2\pi)^{(k+1)/2} \sqrt{2}}.$$

By using (15) and (A.2), we get

$$\lim_{n \rightarrow \infty} \frac{g(\hat{\beta}, \hat{\tau}; \hat{\beta}, \hat{\tau})}{[n/(2\pi)]^{(k+1)/2} |\mathbf{J}_n(\hat{\beta}, \hat{\tau})|^{1/2}} = \lim_{n \rightarrow \infty} \frac{A_{n,k}}{\check{A}_{n,k}}$$

$$\begin{aligned}
&= \lim_{n \rightarrow \infty} \frac{n^{(n-k-1)/2} \sqrt{2\pi}}{2^{(n-k-1)/2} \exp(\frac{n}{2}) \Gamma(\frac{n-k}{2})} \\
&= \lim_{n \rightarrow \infty} \frac{(1-\frac{k}{n})^{(k+1)/2}}{(1-\frac{k/2}{n/2})^{n/2} \exp(\frac{k}{2})} = 1.
\end{aligned} \quad (\text{A.10})$$

The identity in (A.10) was obtained by taking $z = (n-k)/2$ in the well-known expression of the Gamma function:

$$\Gamma(z) = z^{z-1/2} \exp(-z) \exp[\mu(z)] \sqrt{2\pi}, \quad (\text{A.11})$$

where $\mu(z) = \bar{\mu}/(12z)$ and $\bar{\mu} \in (0,1)$. Remark that the Stirling approximation in (26) is a straightforward consequence of (A.11).

Our approach is slightly different than the one from [26, Section 5.2.2] where the condition (A.9) was employed to prove (23). More precisely, we consider the following asymptotic approximation:

$$g(\hat{\beta}, \hat{\tau}; \hat{\beta}, \hat{\tau}) \approx \left(\frac{n}{2\pi}\right)^{(k+1)/2} |\mathbf{J}_{\infty}(\hat{\beta}, \hat{\tau})|^{1/2} = \check{A}_{\infty,k} \hat{\tau}^{-k/2-1}, \quad (\text{A.12})$$

where

$$\begin{aligned}
\check{A}_{\infty,k} &= \left(\frac{n}{2\pi}\right)^{(k+1)/2} \left(\frac{|\mathbf{G}_{\infty}|}{2}\right)^{1/2}, \\
\mathbf{G}_{\infty} &= \lim_{n \rightarrow \infty} \mathbf{G}_n,
\end{aligned} \quad (\text{A.13})$$

$$\mathbf{G}_n = \frac{\mathbf{X}^T \mathbf{X}}{n}. \quad (\text{A.14})$$

Because we want to apply the same techniques like in the evaluation of $\hat{f}_{\rho}(\mathbf{y})$, we use in (A.3) the approximation from (A.12), which leads to

$$C_{\rho}^{\text{FIM}}(R, \tau_0) = (2\check{A}_{\infty,k}/k) \tau_0^{-k/2} V_{\rho}(R). \quad (\text{A.15})$$

It is important to remark that the expression of $\bar{C}_{\rho}(R_1, R_2, \tau_1, \tau_2)$ remains unchanged when (A.5) is modified as follows: (i) $g(\hat{\beta}, \hat{\tau}; \hat{\beta}, \hat{\tau})$ is replaced by the approximation given in (A.12); (ii) $C_{\rho}(\hat{R}, \hat{\tau}_0)$ is replaced by $C_{\rho}^{\text{FIM}}(\hat{R}, \hat{\tau}_0)$. Consequently, the approximate formula of SC is

$$-\ln \hat{f}_{\rho}^{\text{FIM}}(\mathbf{y}) = -\ln f(\mathbf{y}; \hat{\beta}, \hat{\tau}) + \ln C_{\rho}^{\text{FIM}}(\rho(\hat{\beta}), \hat{\tau}) + \ln \bar{C}_{\rho}(R_1, R_2, \tau_1, \tau_2).$$

Furthermore, we compare this result with the one from (20):

$$\begin{aligned}
-\ln \frac{\hat{f}_{\rho}(\mathbf{y})}{\hat{f}_{\rho}^{\text{FIM}}(\mathbf{y})} &= \ln \frac{C_{\rho}(\rho(\hat{\beta}), \hat{\tau})}{C_{\rho}^{\text{FIM}}(\rho(\hat{\beta}), \hat{\tau})} = \ln \frac{A_{n,k}}{\check{A}_{\infty,k}} \\
&= \frac{1}{2} \ln \frac{|\mathbf{G}_n|}{|\mathbf{G}_{\infty}|} + \frac{n-k-1}{2} \ln \frac{n}{2} - \ln \Gamma\left(\frac{n-k}{2}\right) + \frac{1}{2} \ln(2\pi) - \frac{n}{2}
\end{aligned} \quad (\text{A.16})$$

$$\approx \frac{1}{2} \ln \frac{|\mathbf{G}_n|}{|\mathbf{G}_{\infty}|} - \frac{n-k-1}{2} \ln \frac{n-k}{n} - \frac{k}{2}. \quad (\text{A.17})$$

Eq. (A.16) shows clearly that the difference $[-\ln \hat{f}_{\rho}(\mathbf{y})] - [-\ln \hat{f}_{\rho}^{\text{FIM}}(\mathbf{y})]$ does not depend on the constraint $\rho(\cdot)$ which is used for computing the integral. Moreover, based on (A.10), (A.13), (A.14), (A.17), it is easy to conclude that $\hat{f}_{\rho}(\mathbf{y})$ and

$\hat{f}_{\rho}^{\text{FIM}}(\mathbf{y})$ are the same when n is large. Note that the derivation of (A.17) involves the Stirling approximation (26).

Appendix B. Proofs of the main results within Section 3.2

Proof of Proposition 3.1.

(a) We consider the singular value decomposition (SVD) of the matrix \mathbf{X}_{γ} . Let $\mathbf{X}_{\gamma} = [\mathbf{U} \mathbf{U}_0][\mathbf{\Lambda}^T \mathbf{0}^T]^T \mathbf{V}^T$, where the matrix $[\mathbf{U} \mathbf{U}_0]$ has orthonormal columns, $\mathbf{U} \in \mathbb{R}^{n \times k}$ and $\mathbf{U}_0 \in \mathbb{R}^{n \times (n-k)}$. The diagonal matrix $\mathbf{\Lambda} \in \mathbb{R}^{k \times k}$ is non-singular, and $\mathbf{V} \in \mathbb{R}^{k \times k}$ is such that $\mathbf{V}^{-1} = \mathbf{V}^T$. For $i \in \{1, 2, 3\}$, we have $\mathbf{Q}_i = \mathbf{V} \mathbf{L}_i^2 \mathbf{V}^T$ and

$$D_{\gamma}(\mathbf{y}; \mathbf{Q}_i) = \mathbf{y}^T \mathbf{U} \frac{\mathbf{L}_i^2}{|\mathbf{L}_i^2|^{1/k}} \mathbf{M}^{-1} \mathbf{U}^T \mathbf{y},$$

where $\mathbf{L}_1 = \mathbf{\Lambda}$, $\mathbf{L}_2 = \mathbf{I}$, $\mathbf{L}_3 = \mathbf{\Lambda}^2$ and $\mathbf{M} = \mathbf{\Lambda}^2 / |\mathbf{\Lambda}^2|^{1/k}$. The equalities in (35) can be re-written as

$$\|\mathbf{U}^T \mathbf{y}\|^2 = \|\mathbf{M}^{-1/2} \mathbf{U}^T \mathbf{y}\|^2 = \|\mathbf{M}^{1/2} \mathbf{U}^T \mathbf{y}\|^2,$$

and they are satisfied for all $\mathbf{y} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ if and only if $\mathbf{M} = \mathbf{I}$. This is equivalent with the fact that $\mathbf{\Lambda}^2$ has one eigenvalue with multiplicity k . We denote q this eigenvalue, and the condition in (36) is immediately obtained.

(b) We use the notations introduced in the proof of the point (a), and we focus on the properties of the matrix \mathbf{M} . Observe that the diagonal entries of \mathbf{M} are strictly positive, and their product is equal to one. If $\mathbf{M} \neq \mathbf{I}$, then some of the eigenvalues of \mathbf{M}^{-1} are larger than one, while the others are smaller than one. Therefore, the matrix $\mathbf{I} - \mathbf{M}^{-1}$ has both positive and negative eigenvalues. This observation together with the identity $D_{\gamma}(\mathbf{y}; \mathbf{Q}_1) - D_{\gamma}(\mathbf{y}; \mathbf{Q}_2) = \mathbf{y}^T \mathbf{U} (\mathbf{I} - \mathbf{M}^{-1}) \mathbf{U}^T \mathbf{y}$ show that, depending on \mathbf{y} , the difference $D_{\gamma}(\mathbf{y}; \mathbf{Q}_1) - D_{\gamma}(\mathbf{y}; \mathbf{Q}_2)$ can be either positive or negative. The proof is similar for $D_{\gamma}(\mathbf{y}; \mathbf{Q}_2) - D_{\gamma}(\mathbf{y}; \mathbf{Q}_3)$ and $D_{\gamma}(\mathbf{y}; \mathbf{Q}_1) - D_{\gamma}(\mathbf{y}; \mathbf{Q}_3)$.

(c) It is easy to verify that $D_{\gamma}(\mathbf{y}; \mathbf{Q}_2) \times D_{\gamma}(\mathbf{y}; \mathbf{Q}_3) = \|\hat{\beta}_{\gamma}\|^2 \times \|\mathbf{X}_{\gamma}^T \mathbf{y}\|^2$ and $D_{\gamma}(\mathbf{y}; \mathbf{Q}_1)^2 = [\hat{\beta}_{\gamma}^T (\mathbf{X}_{\gamma}^T \mathbf{y})]^2$. The Cauchy-Schwarz inequality [30, p. 258] written for the vectors $\hat{\beta}_{\gamma}$ and $\mathbf{X}_{\gamma}^T \mathbf{y}$ leads to $D_{\gamma}(\mathbf{y}; \mathbf{Q}_2) \times D_{\gamma}(\mathbf{y}; \mathbf{Q}_3) \geq D_{\gamma}(\mathbf{y}; \mathbf{Q}_1)^2$, which proves the inequality in (37). \square

Proof of Proposition 3.2. The main idea is to write the expressions of $D_{\gamma}(\mathbf{y}; \mathbf{Q}_i)$, $i \in \{1, 2, 3\}$, in a form which allows us to see immediately if, for $\alpha \rightarrow 0$, the result is finite or not. We introduce the following supplementary notations: $\mathbf{g} = \mathbf{P}_{k-1}^{\perp} \mathbf{x}_k$, $\mathbf{P}_k = \mathbf{X}_k (\mathbf{X}_k^T \mathbf{X}_k)^{-1} \mathbf{X}_k^T$, $\mathbf{P}_{\mathbf{x}_k} = (\mathbf{x}_k \mathbf{x}_k^T) / \|\mathbf{x}_k\|^2$ and $\mathbf{P}_{\mathbf{x}_k}^{\perp} = \mathbf{I} - \mathbf{P}_{\mathbf{x}_k}$. The symbol $\#$ is used for the Moore-Penrose pseudoinverse.

(a) Some simple manipulations combined with the identity from [14, Eq. (8.34)] lead to

$$\begin{aligned}
D_{\gamma}(\mathbf{y}; \mathbf{Q}_1) &= \|\mathbf{P}_k \mathbf{y}\|^2 = \left\| \left(\mathbf{P}_{k-1} + \frac{\mathbf{g} \mathbf{g}^T}{\|\mathbf{g}\| \|\mathbf{g}\|} \right) \mathbf{y} \right\|^2 \\
&= \|(\mathbf{P}_{k-1} + \mathbf{w} \mathbf{w}^T) \mathbf{y}\|^2.
\end{aligned} \quad (\text{B.1})$$

- (b) To compute $D_\gamma(\mathbf{y}; \mathbf{Q}_2)$, we use the formula [20, Eq. (2.17)]:

$$(\mathbf{X}_k^\top \mathbf{X}_k)^{-1} = \begin{bmatrix} (\mathbf{X}_{k-1}^\top \mathbf{P}_{\mathbf{x}_k}^\perp \mathbf{X}_{k-1})^{-1} & \mathbf{F} \\ \mathbf{F}^\top & (\mathbf{x}_k^\top \mathbf{P}_{k-1}^\perp \mathbf{x}_k)^{-1} \end{bmatrix}, \quad (\text{B.2})$$

where $\mathbf{F} = -\mathbf{X}_{k-1}^\top \mathbf{x}_k (\mathbf{x}_k^\top \mathbf{P}_{k-1}^\perp \mathbf{x}_k)^{-1}$. Simple calculations produce the following outcome:

$$(\mathbf{X}_k^\top \mathbf{X}_k)^{-1} \mathbf{X}_k^\top = \begin{bmatrix} (\mathbf{P}_{\mathbf{x}_k}^\perp \mathbf{X}_{k-1})^\# \\ \frac{1}{\|\mathbf{g}\|} \frac{\mathbf{g}^\top}{\|\mathbf{g}\|} \end{bmatrix}. \quad (\text{B.3})$$

Then we employ the identity from [2, Eq. (17)] to get

$$\begin{aligned} (\mathbf{P}_{\mathbf{x}_k}^\perp \mathbf{X}_{k-1})^\# &= (\mathbf{X}_{k-1}^\top \mathbf{P}_{\mathbf{x}_k}^\perp \mathbf{X}_{k-1})^{-1} \mathbf{X}_{k-1}^\top \mathbf{P}_{\mathbf{x}_k}^\perp \\ &= (\mathbf{X}_{k-1}^\top \mathbf{X}_{k-1}) (\mathbf{X}_{k-1}^\top \mathbf{P}_{\mathbf{x}_k}^\perp \mathbf{X}_{k-1})^{-1} \mathbf{X}_{k-1}^\top \mathbf{P}_{\mathbf{x}_k}^\perp \\ &= \mathbf{X}_{k-1}^\# \mathbf{P}_{k-1} [\mathbf{I} - \mathbf{x}_k (\mathbf{x}_k^\top \mathbf{P}_{k-1}^\perp \mathbf{x}_k)^{-1} \mathbf{x}_k^\top \mathbf{P}_{k-1}^\perp] \\ &= \mathbf{X}_{k-1}^\# \left(\|\mathbf{g}\| \mathbf{I} - \mathbf{x}_k \frac{\mathbf{g}^\top}{\|\mathbf{g}\|} \right). \end{aligned}$$

The result above together with (B.3) show that

$$\begin{aligned} \mathbf{X}_k (\mathbf{X}_k^\top \mathbf{X}_k)^{-2} \mathbf{X}_k^\top &= \frac{1}{\|\mathbf{g}\|^2} \left(\|\mathbf{g}\| \mathbf{I} - \mathbf{x}_k \frac{\mathbf{g}^\top}{\|\mathbf{g}\|} \right)^\top (\mathbf{X}_{k-1}^\#)^\top \\ &\quad \times \mathbf{X}_{k-1}^\# \left(\|\mathbf{g}\| \mathbf{I} - \mathbf{x}_k \frac{\mathbf{g}^\top}{\|\mathbf{g}\|} \right) + \frac{1}{\|\mathbf{g}\|^2} \frac{\mathbf{g} \mathbf{g}^\top}{\|\mathbf{g}\|}. \end{aligned} \quad (\text{B.4})$$

Additionally, it is known that [16]

$$|\mathbf{X}_k^\top \mathbf{X}_k| = \|\mathbf{g}\|^2 |\mathbf{X}_{k-1}^\top \mathbf{X}_{k-1}|. \quad (\text{B.5})$$

So,

$$\begin{aligned} D_\gamma(\mathbf{y}; \mathbf{Q}_2) &= \frac{|\mathbf{X}_{k-1}^\top \mathbf{X}_{k-1}|^{1/k}}{\|\mathbf{g}\|^{2(1-1/k)}} \left[\left\| \mathbf{X}_{k-1}^\# \left(\|\mathbf{g}\| \mathbf{I} - \mathbf{x}_k \frac{\mathbf{g}^\top}{\|\mathbf{g}\|} \right) \mathbf{y} \right\|^2 \right. \\ &\quad \left. + \left(\frac{\mathbf{g}^\top}{\|\mathbf{g}\|} \mathbf{y} \right)^2 \right] = \frac{|\mathbf{X}_{k-1}^\top \mathbf{X}_{k-1}|^{1/k}}{[\sin(\alpha) \|\mathbf{x}_k\|]^{2(1-1/k)}} \\ &\quad \times [|\mathbf{X}_{k-1}^\# (\sin(\alpha) \|\mathbf{x}_k\| \mathbf{I} - \mathbf{x}_k \mathbf{w}^\top) \mathbf{y}|^2 + (\mathbf{w}^\top \mathbf{y})^2]. \end{aligned} \quad (\text{B.6})$$

- (c) It is obvious that $D_\gamma(\mathbf{y}; \mathbf{Q}_3) = \|\mathbf{X}_k^\top \mathbf{y}\|^2 / |\mathbf{X}_k^\top \mathbf{X}_k|^{1/k}$. Then we apply (B.5) to get

$$\begin{aligned} D_\gamma(\mathbf{y}; \mathbf{Q}_3) &= \frac{1}{\|\mathbf{g}\|^{2/k}} \frac{\|\mathbf{X}_k^\top \mathbf{y}\|^2}{|\mathbf{X}_{k-1}^\top \mathbf{X}_{k-1}|^{1/k}} \\ &= \frac{1}{[\sin(\alpha) \|\mathbf{x}_k\|]^{2/k}} \frac{\|\mathbf{X}_k^\top \mathbf{y}\|^2}{|\mathbf{X}_{k-1}^\top \mathbf{X}_{k-1}|^{1/k}}. \end{aligned} \quad (\text{B.7})$$

Proposition 3.2 is a straightforward consequence of (B.1), (B.6) and (B.7). \square

Proof of Proposition 3.3. The equality in (41) is readily obtained from (34) and (39). We also have from (34) that

$$D_\gamma(\mathbf{y}; \mathbf{Q}_2) = \|\hat{\boldsymbol{\beta}}_\gamma\|^2 \times |\mathbf{X}_\gamma^\top \mathbf{X}_\gamma|^{1/k}. \quad (\text{B.8})$$

Let $\hat{\beta}_k$ be the last entry of the vector $\hat{\boldsymbol{\beta}}_\gamma$. With the notations from the proof of Proposition 3.2, we have

$$\hat{\beta}_k^2 = \left(\frac{\mathbf{g}^\top \mathbf{y}}{\|\mathbf{g}\|^2} \right)^2 \quad (\text{B.9})$$

$$\begin{aligned} &= \frac{\mathbf{y}^\top [(\mathbf{g}\mathbf{g}^\top) / \|\mathbf{g}\|^2] \mathbf{y}}{\|\mathbf{g}\|^2} = \frac{\mathbf{y}^\top (\mathbf{P}_k - \mathbf{P}_{k-1}) \mathbf{y}}{\|\mathbf{P}_k^\perp \mathbf{x}_k\|^2} \\ &= \frac{\|\mathbf{P}_k \mathbf{y}\|^2 - \|\mathbf{P}_{k-1} \mathbf{y}\|^2}{1 - \|\mathbf{P}_{k-1} \mathbf{x}_k\|^2} = \frac{R_{\mathbf{y}\mathbf{x}_\gamma}^2 - R_{\mathbf{y}\mathbf{x}_{\gamma(i)}}^2}{1 - R_{k-1, \dots, (k-1)}^2}. \end{aligned} \quad (\text{B.10})$$

Note that (B.9) is obtained from (B.3), and (B.10) is based on (B.1). The result can be extended to all entries of $\hat{\boldsymbol{\beta}}_\gamma$, and we get

$$\|\hat{\boldsymbol{\beta}}_\gamma\|^2 = \sum_{i=1}^k \frac{R_{\mathbf{y}\mathbf{x}_\gamma}^2 - R_{\mathbf{y}\mathbf{x}_{\gamma(i)}}^2}{1 - R_{\zeta(k), \zeta(1), \dots, \zeta(k-1)}^2}, \quad (\text{B.11})$$

where $\zeta(\cdot)$ is defined in (44). Next we use recursively the identity from (B.5) to obtain

$$|\mathbf{X}_\gamma^\top \mathbf{X}_\gamma| = \prod_{i=2}^k \|\mathbf{P}_{i-1}^\perp \mathbf{x}_i\|^2 = \prod_{i=2}^k [1 - R_{i-1, \dots, (i-1)}^2]. \quad (\text{B.12})$$

For an arbitrary $i \in \{1, \dots, k-1\}$, we consider the matrix $\tilde{\mathbf{X}}_\gamma = [\mathbf{x}_1 \cdots \mathbf{x}_{i-1} \mathbf{x}_{i+1} \cdots \mathbf{x}_k \mathbf{x}_i]$, which is obtained by permuting the columns of \mathbf{X}_γ . For computing $|\tilde{\mathbf{X}}_\gamma^\top \tilde{\mathbf{X}}_\gamma|$, we apply the same technique like in (B.12). The fact that $|\tilde{\mathbf{X}}_\gamma^\top \tilde{\mathbf{X}}_\gamma| = |\mathbf{X}_\gamma^\top \mathbf{X}_\gamma|$ leads to

$$|\mathbf{X}_\gamma^\top \mathbf{X}_\gamma| = [1 - R_{\zeta(k), \zeta(1), \dots, \zeta(k-1)}^2] \prod_{j=2}^{k-1} [1 - R_{\zeta(j), \zeta(1), \dots, \zeta(j-1)}^2]. \quad (\text{B.13})$$

The identity in (42) is proven by combining (B.8), (B.11) and (B.13). We notice from (34) that $D_\gamma(\mathbf{y}; \mathbf{Q}_3) = \|\mathbf{X}_\gamma^\top \mathbf{y}\|^2 / |\mathbf{X}_\gamma^\top \mathbf{X}_\gamma|^{1/k}$, and by using (B.12), we get (43). \square

Appendix C. Proofs of the main results within Section 3.3

Proof of Lemma 3.1. For $i \in \{1, 2, 3\}$, we define $\mathbf{M}_i = \mathbf{X}_k (\mathbf{X}_k^\top \mathbf{X}_k)^{-1} \mathbf{Q}_i (\mathbf{X}_k^\top \mathbf{X}_k)^{-1} \mathbf{x}_k^\top / (|\mathbf{Q}_i| / |\mathbf{X}_k^\top \mathbf{X}_k|)^{1/k}$, and by applying a well-known result [30, p. 439], we have

$$\begin{aligned} \mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_i)] &= \mathbb{E}[\mathbf{y}^\top] \mathbf{M}_i \mathbb{E}[\mathbf{y}] + \tau \text{Tr}[\mathbf{M}_i] \\ &= \boldsymbol{\beta}_{k-1}^\top \mathbf{X}_{k-1}^\top \mathbf{M}_i \mathbf{X}_{k-1} \boldsymbol{\beta}_{k-1} + \tau \frac{\text{Tr}[(\mathbf{X}_k^\top \mathbf{X}_k)^{-1} \mathbf{Q}_i]}{(|\mathbf{Q}_i| / |\mathbf{X}_k^\top \mathbf{X}_k|)^{1/k}}, \end{aligned} \quad (\text{C.1})$$

where $\text{Tr}[\cdot]$ denotes the trace operator. When $\mathbf{Q} = \mathbf{Q}_1$, we compute (C.1) by making use of techniques similar with those employed to derive (B.1):

$$\begin{aligned} \mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)] &= \|\mathbf{P}_k \mathbf{X}_{k-1} \boldsymbol{\beta}_{k-1}\|^2 + \tau \text{Tr}[\mathbf{I}] \\ &= \|(\mathbf{P}_{k-1} + \omega^{-1} \mathbf{P}_{k-1}^\perp \mathbf{x}_k \mathbf{x}_k^\top \mathbf{P}_{k-1}^\perp) \mathbf{X}_{k-1} \boldsymbol{\beta}_{k-1}\|^2 + \tau k \\ &= \|\boldsymbol{\beta}_{k-1}\|^2 + \tau k. \end{aligned}$$

Hence, the identity in (45) is proven. Next we focus on some results that will be useful when evaluating (C.1) for $\mathbf{Q} = \mathbf{Q}_2$ and $\mathbf{Q} = \mathbf{Q}_3$. First notice from (B.5) that $|\mathbf{X}_k^\top \mathbf{X}_k| = \omega$. Moreover, we have from (B.2) that

$$\begin{aligned} \text{Tr}[(\mathbf{X}_k^\top \mathbf{X}_k)^{-1}] &= \text{Tr}[(\mathbf{X}_{k-1}^\top \mathbf{P}_{\mathbf{x}_k}^\perp \mathbf{X}_{k-1})^{-1}] + (\mathbf{x}_k^\top \mathbf{P}_{k-1}^\perp \mathbf{x}_k)^{-1} \\ &= \text{Tr}[(\mathbf{X}_{k-1}^\top \mathbf{P}_{\mathbf{x}_k}^\perp \mathbf{X}_{k-1})^{-1}] + \omega^{-1} = k - 2 + 2\omega^{-1}. \end{aligned} \quad (\text{C.2})$$

The identity above is deduced by taking into account that the eigenvalues of $(\mathbf{X}_{k-1}^\top \mathbf{P}_{\mathbf{x}_k}^\perp \mathbf{X}_{k-1})^{-1}$ are 1 and ω^{-1} . The

eigenvalue 1 has multiplicity $k-2$, while the eigenvalue ω^{-1} has multiplicity 1 [2, Eq. (25)]. These results together with (C.1) and some algebra yield (46) and (47):

$$\begin{aligned}\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_2)] &= [\|\mathbf{X}_k^\top \mathbf{X}_{k-1} \boldsymbol{\beta}_{k-1}\|^2 + \tau \text{Tr}[(\mathbf{X}_k^\top \mathbf{X}_k)^{-1}] \|\mathbf{X}_k^\top \mathbf{X}_k\|^{1/k}] \\ &= [\|\boldsymbol{\beta}_{k-1}\|^2 + \tau(k-2+2\omega^{-1})] \omega^{1/k},\end{aligned}$$

$$\begin{aligned}\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_3)] &= [\|\mathbf{X}_k^\top \mathbf{X}_{k-1} \boldsymbol{\beta}_{k-1}\|^2 + \tau \text{Tr}[(\mathbf{X}_k^\top \mathbf{X}_k)^{-1}] \|\mathbf{X}_k^\top \mathbf{X}_k\|^{1/k}] \\ &= [\|\boldsymbol{\beta}_{k-1}\|^2 + \tau k + (\mathbf{x}_k^\top \mathbf{X}_{k-1} \boldsymbol{\beta}_{k-1})^2] \omega^{-1/k}.\end{aligned}\quad \square$$

Proof of Proposition 3.4.

(a) It follows from (45) and (46) that

$$\begin{aligned}\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_2) - D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)] &= \|\boldsymbol{\beta}_{k-1}\|^2 [\omega^{1/k} - 1] + \tau[k(\omega^{1/k} - 1) - 2\omega^{1/k-1}(\omega - 1)] \\ &= \tau(\omega^{1/k} - 1) \left[\frac{\|\boldsymbol{\beta}_{k-1}\|^2}{\tau} + k - 2\omega^{1/k-1} \frac{\omega - 1}{\omega^{1/k} - 1} \right] \\ &= \tau(\omega^{1/k} - 1) \left\{ \frac{\|\boldsymbol{\beta}_{k-1}\|^2}{\tau} - \left[2 \frac{1 - \omega^{-1}}{1 - \omega^{-1/k}} - k \right] \right\}.\end{aligned}$$

We can now infer the conclusion within point (a) of Proposition 3.4 by noticing that $\omega^{1/k} - 1 < 0$ for $\alpha \in (0, \pi/2)$ and, additionally, $2(1 - \omega^{-1})/(1 - \omega^{-1/k}) - k$ is a decreasing function of α .

(b) The identities in (45) and (47) prove that

$$\begin{aligned}\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_3) - D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)] &= \mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)](\omega^{-1/k} - 1) + (\mathbf{x}_k^\top \mathbf{X}_{k-1} \boldsymbol{\beta}_{k-1})^2 \omega^{-1/k} \\ &\leq (\|\boldsymbol{\beta}_{k-1}\|^2 + \tau k)(\omega^{-1/k} - 1) + \|\boldsymbol{\beta}_{k-1}\|^2 (1 - \omega) \omega^{-1/k}\end{aligned}\quad (\text{C.3})$$

$$\leq (\|\boldsymbol{\beta}_{k-1}\|^2 + \tau k)(\omega^{-1/k} - 1) + \|\boldsymbol{\beta}_{k-1}\|^2 (1 - \omega) \omega^{-1/k} \quad (\text{C.4})$$

It is evident from (C.3) that $\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_3) - D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)]$ cannot be negative because both $\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)](\omega^{-1/k} - 1)$ and $(\mathbf{x}_k^\top \mathbf{X}_{k-1} \boldsymbol{\beta}_{k-1})^2 \omega^{-1/k}$ are non-negative. Note that (C.4) is obtained by applying the Cauchy-Schwarz inequality [30, p. 258]:

$$(\mathbf{x}_k^\top \mathbf{X}_{k-1} \boldsymbol{\beta}_{k-1})^2 \leq \|\mathbf{x}_k^\top \mathbf{X}_{k-1}\|^2 \|\boldsymbol{\beta}_{k-1}\|^2 = (1 - \omega) \|\boldsymbol{\beta}_{k-1}\|^2.$$

The inequality in (48) is a straightforward consequence of (C.4). \square

Proof of Proposition 3.5. First we give three auxiliary results which will be instrumental for the main proof.

- **Result #1** [36, p. 348]. Let $\mathbf{N} \in \mathbb{R}^{k \times k}$ be a symmetric matrix whose eigenvalues are v_1, \dots, v_k . Also, let $\boldsymbol{\theta} \in \mathbb{R}^k \setminus \{\mathbf{0}\}$. Then

$$v_{\min} \|\boldsymbol{\theta}\|^2 \leq \boldsymbol{\theta}^\top \mathbf{N} \boldsymbol{\theta} \leq v_{\max} \|\boldsymbol{\theta}\|^2, \quad (\text{C.5})$$

where $v_{\min} = \min_{1 \leq i \leq k} v_i$ and $v_{\max} = \max_{1 \leq i \leq k} v_i$.

- **Result #2.** The arithmetic-geometric-harmonic mean inequalities [21, p. 27] applied to the eigenvalues of $\mathbf{X}_k^\top \mathbf{X}_k$:

$$\lambda_{\min} \leq \mathcal{H}_\lambda \leq \mathcal{G}_\lambda \leq \mathcal{A}_\lambda \leq \lambda_{\max}, \quad (\text{C.6})$$

where $\lambda_{\min} = \min_{1 \leq i \leq k} \lambda_i$ and $\lambda_{\max} = \max_{1 \leq i \leq k} \lambda_i$.

- **Result #3.** If $\mathbf{X}_{k-1}^\top \mathbf{X}_{k-1} = \mathbf{I}$, $\|\mathbf{x}_k\| = 1$ and $\alpha \in (0, \pi/2)$ is the principal angle between $\langle \mathbf{x}_{k-1} \rangle$ and $\langle \mathbf{x}_k \rangle$, then the

eigenvalues of $\mathbf{X}_k^\top \mathbf{X}_k$ satisfy the inequalities

$$1 - \cos \alpha \leq \lambda_{\min} \leq \lambda_{\max} \leq 1 + \cos \alpha. \quad (\text{C.7})$$

Proof: Let $\mathbf{b} = \mathbf{X}_{k-1}^\top \mathbf{x}_k$ and $\mathbf{B} = \begin{bmatrix} \mathbf{0} & \mathbf{b} \\ \mathbf{b}^\top & 0 \end{bmatrix}$. The equality $\mathbf{B} = \mathbf{X}_k^\top \mathbf{X}_k - \mathbf{I}$ is evident, and it implies that the eigenvalues of \mathbf{B} are $\lambda_1 - 1, \dots, \lambda_k - 1$. For $i \in \{1, \dots, k\}$, if \mathbf{v}_i is the eigenvector of $\mathbf{X}_k^\top \mathbf{X}_k$ associated with λ_i , then \mathbf{v}_i is also the eigenvector of \mathbf{B} associated with $\lambda_i - 1$. With the convention that $\mathbf{b} = [b_1, \dots, b_{k-1}]^\top$ and $\mathbf{v}_i = [v_{1,i}, \dots, v_{k,i}]^\top$, we have

$$(\lambda_i - 1) \mathbf{v}_i = \begin{bmatrix} \mathbf{0} & \mathbf{b} \\ \mathbf{b}^\top & 0 \end{bmatrix} \mathbf{v}_i = \begin{bmatrix} b_1 v_{k,i} \\ \vdots \\ b_{k-1} v_{k,i} \\ \sum_{j=1}^{k-1} b_j v_{j,i} \end{bmatrix}.$$

The identities $\|\mathbf{v}_i\|^2 = 1$ and $\|\mathbf{b}\|^2 = \cos^2 \alpha$ together with the Cauchy-Schwarz inequality [30, p. 258] yield

$$\begin{aligned}(\lambda_i - 1)^2 &= v_{k,i}^2 \|\mathbf{b}\|^2 + \left(\sum_{j=1}^{k-1} b_j v_{j,i} \right)^2 \\ &\leq v_{k,i}^2 \|\mathbf{b}\|^2 + \|\mathbf{b}\|^2 \sum_{j=1}^{k-1} v_{j,i}^2 = \cos^2 \alpha,\end{aligned}$$

which implies $1 - \cos \alpha \leq \lambda_i \leq 1 + \cos \alpha$ for all $i \in \{1, \dots, k\}$. \square

Main inequalities:

(a) From (49) and (50), we get

$$\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_2) - D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)] = \boldsymbol{\beta}_k^\top \mathbf{L} \boldsymbol{\beta}_k + \tau k(\mathcal{G}_\lambda / \mathcal{H}_\lambda - 1),$$

where $\mathbf{L} = \mathcal{G}_\lambda \mathbf{I} - \mathbf{X}_k^\top \mathbf{X}_k$. Observe that the smallest eigenvalue of \mathbf{L} is

$$\ell_{\min} = \mathcal{G}_\lambda - \lambda_{\max}, \quad (\text{C.8})$$

and the largest eigenvalue of \mathbf{L} is

$$\ell_{\max} = \mathcal{G}_\lambda - \lambda_{\min}. \quad (\text{C.9})$$

By making use of (B.5), it is easy to check that

$$\mathcal{G}_\lambda = \sin^{2/k} \alpha. \quad (\text{C.10})$$

The steps of the proof for the inequalities in (52) are outlined below. At each step, we indicate which result is used in demonstration.

$$\begin{aligned}\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_2) - D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)] &\stackrel{(\text{C.6})}{\geq} \boldsymbol{\beta}_k^\top \mathbf{L} \boldsymbol{\beta}_k \\ &\stackrel{(\text{C.5})}{\geq} \|\boldsymbol{\beta}_k\|^2 \ell_{\min} \\ &\stackrel{(\text{C.8})}{=} \|\boldsymbol{\beta}_k\|^2 (\mathcal{G}_\lambda - \lambda_{\max}) \\ &\stackrel{(\text{C.10})}{=} \|\boldsymbol{\beta}_k\|^2 (\sin^{2/k} \alpha - \lambda_{\max}) \\ &\stackrel{(\text{C.7})}{\geq} \|\boldsymbol{\beta}_k\|^2 (\sin^{2/k} \alpha - \cos \alpha - 1),\end{aligned}$$

$$\begin{aligned}\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_2) - D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)] &\stackrel{(\text{C.6})}{\leq} \boldsymbol{\beta}_k^\top \mathbf{L} \boldsymbol{\beta}_k + \tau k(\mathcal{G}_\lambda / \lambda_{\min} - 1) \\ &\stackrel{(\text{C.5})}{\leq} \|\boldsymbol{\beta}_k\|^2 \ell_{\max} + \tau k(\mathcal{G}_\lambda / \lambda_{\min} - 1) \\ &\stackrel{(\text{C.9})}{=} \|\boldsymbol{\beta}_k\|^2 (\mathcal{G}_\lambda - \lambda_{\min}) + \tau k(\mathcal{G}_\lambda / \lambda_{\min} - 1) \\ &\stackrel{(\text{C.7})}{\leq} \|\boldsymbol{\beta}_k\|^2 (\mathcal{G}_\lambda + \cos \alpha - 1) + \tau k[\mathcal{G}_\lambda / (1 - \cos \alpha) - 1]\end{aligned}$$

$$\stackrel{(C.10)}{=} \|\beta_k\|^2 (\sin^{2/k} \alpha + \cos \alpha - 1) + \tau k \left(\frac{\sin^{2/k} \alpha}{1 - \cos \alpha} - 1 \right).$$

(b) By subtracting (49) from (51), we obtain

$$\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_3) - D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)] = \beta_k^\top \mathbf{M} \beta_k + \tau k (\mathcal{A}_k / \mathcal{G}_k - 1),$$

where $\mathbf{M} = (\mathbf{X}_k^\top \mathbf{X}_k)^2 / \mathcal{G}_k - \mathbf{X}_k^\top \mathbf{X}_k$. Let us consider the mapping $\phi(z) = z^2 / \mathcal{G}_k - z$, which is defined for all $z \in \mathbb{R}$. The eigenvalues of \mathbf{M} are $\mu_i = \phi(\lambda_i)$, $i \in \{1, \dots, k\}$. The inequalities in (C.7), together with the well-known properties of $\phi(\cdot)$, guarantee that μ_{\max} , the maximum eigenvalue of \mathbf{M} , has the property: $\mu_{\max} \leq \max\{\phi(1 - \cos \alpha), \phi(1 + \cos \alpha)\}$. Because $\phi(1 + \cos \alpha) - \phi(1 - \cos \alpha) = 2 \cos \alpha (2 \sin^{-2/k} \alpha - 1) > 0$, the following inequality holds true:

$$\mu_{\max} \leq \phi(1 + \cos \alpha). \quad (C.11)$$

Since the parabola defined by $\phi(\cdot)$ attains its minimum when $z = \mathcal{G}_k / 2$, it is obvious that μ_{\min} , the minimum eigenvalue of \mathbf{M} , cannot be smaller than $\phi(\mathcal{G}_k / 2)$. So,

$$\mu_{\min} \geq \phi(\mathcal{G}_k / 2). \quad (C.12)$$

The inequality above can be improved by observing for $\alpha \geq \pi/3$ that $1 - \cos \alpha \geq \mathcal{G}_k / 2$ for all $k \geq 2$. In this case,

$$\mu_{\min} \geq \phi(1 - \cos \alpha). \quad (C.13)$$

Similarly with the chain of inequalities for $\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_2) - D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)]$, we write

$$\mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_3) - D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)] \stackrel{(C.6)}{\geq} \beta_k^\top \mathbf{M} \beta_k \stackrel{(C.5)}{\geq} \|\beta_k\|^2 \mu_{\min}.$$

From the inequality above we get (53) and (54) by employing (C.12) and (C.13). Then we focus on the proof of (55):

$$\begin{aligned} \mathbb{E}[D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_3) - D_{\gamma_k}(\mathbf{y}; \mathbf{Q}_1)] &\stackrel{(C.6)}{\leq} \beta_k^\top \mathbf{M} \beta_k + \tau k (\lambda_{\max} / \mathcal{G}_k - 1) \\ &\stackrel{(C.5)}{\leq} \|\beta_k\|^2 \mu_{\max} + \tau k (\lambda_{\max} / \mathcal{G}_k - 1) \\ &\stackrel{(C.7)}{\leq} \|\beta_k\|^2 \mu_{\max} + \tau k [(1 + \cos \alpha) / \mathcal{G}_k - 1] \\ &\stackrel{(C.11)}{\leq} \|\beta_k\|^2 \phi(1 + \cos \alpha) + \tau k [(1 + \cos \alpha) / \mathcal{G}_k - 1] \\ &\stackrel{(C.10)}{=} \|\beta_k\|^2 \left[\frac{(1 + \cos \alpha)^2}{\sin^{2/k} \alpha} - (1 + \cos \alpha) \right] + \tau k \left(\frac{1 + \cos \alpha}{\sin^{2/k} \alpha} - 1 \right). \quad \square \end{aligned}$$

References

- [1] H. Akaike, A new look at the statistical model identification, *IEEE Transactions on Automatic Control* AC-19 (1974) 716–723.
- [2] R. Behrens, L. Scharf, Signal processing applications of oblique projection operators, *IEEE Transactions on Signal Processing* 42 (1994) 1413–1424.
- [3] A. Björck, G. Golub, Numerical methods for computing angles between linear subspaces, *Mathematics of Computation* 27 (1973) 579–594.
- [4] L. Breiman, D. Freedman, How many variables should be entered in a regression equation?, *Journal of the American Statistical Association* 78 (1983) 131–136.
- [5] P. Djuric, Asymptotic MAP criteria for model selection, *IEEE Transactions on Signal Processing* 46 (1998) 2726–2735.
- [6] C. Giurcăneanu, Estimation of sinusoidal regression model by stochastic complexity, in: P. Grünwald, P. Myllymäki, I. Tabus, M. Weinberger, B. Yu (Eds.), *Festschrift in Honor of Jorma Rissanen on the Occasion of his 75th Birthday*, TICSP Series, vol. 38, Tampere International Center for Signal Processing, Tampere, Finland, <http://www.cs.tut.fi/~tabus/TICSP_38_17.4.08.pdf>, 2008, pp. 229–249.
- [7] C. Giurcăneanu, S. Razavi, New insights on stochastic complexity, in: *Proceedings of the Eusipco 2009, the 17th European Signal Processing Conference*, Glasgow, Scotland, UK, pp. 2475–2479.
- [8] P. Grünwald, *The Minimum Description Length Principle*, MIT Press, 2007.
- [9] M. Hansen, B. Yu, Model selection and the principle of minimum description length, *Journal of the American Statistical Association* 96 (2001) 746–774.
- [10] M. Hansen, B. Yu, Minimum description length model selection criteria for generalized linear models, in: D. Goldstein (Ed.), *Science and Statistics: A Festschrift for Terry Speed*, Institute of Mathematical Statistics, Lecture Notes-Monograph Series, vol. 40, 2002, pp. 145–164.
- [11] A. Hanson, P.C.W. Fu, Applications of MDL to selected families of models, in: P. Grünwald, I. Myung, M. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*, MIT Press, 2005, pp. 125–150.
- [12] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, second ed., Springer, 2009.
- [13] C. Hurvich, C.L. Tsai, Regression and time series model selection in small samples, *Biometrika* 76 (1989) 297–307.
- [14] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, 1993.
- [15] S. Kay, Conditional model order estimation, *IEEE Transactions on Signal Processing* 49 (2001) 1910–1917.
- [16] S. Kay, Exponentially embedded families—new approaches to model order estimation, *IEEE Transactions on Aerospace and Electronic Systems* 41 (2005) 333–345.
- [17] E. Liski, Normalized ML and the MDL principle for variable selection in linear regression, in: E. Liski, J. Isotalo, J. Niemelä, S. Puntanen, G. Styan (Eds.), *Festschrift for Tarmo Pukkila on his 60th Birthday*, University of Tampere, 2006, pp. 159–172.
- [18] E. Liski, A. Liski, Minimum description length model selection in Gaussian regression under data constraints, in: B. Schipp, W. Krämer (Eds.), *Statistical Inference, Econometric Analysis and Matrix Algebra*, *Festschrift in Honour of Götz Trenkler*, Springer, 2009, pp. 201–208 <http://www.springerlink.com/content/n32t4671713w87g0/fulltext.pdf>.
- [19] K. Mardia, J. Kent, J. Bibby, *Multivariate Analysis*, Academic Press, 1979.
- [20] T. McWhorter, L. Scharf, Cramer–Rao bounds for deterministic modal analysis, *IEEE Transactions on Signal Processing* 41 (1993) 1847–1866.
- [21] D. Mitrinovic, P.M. Vasic, *Analytic Inequalities*, Springer Verlag, 1970.
- [22] G. Qian, H. Künsch, Some notes on Rissanen’s stochastic complexity, *IEEE Transactions on Information Theory* 44 (1998) 782–786.
- [23] J. Rissanen, Modeling by shortest data description, *Automatica* 14 (1978) 465–471.
- [24] J. Rissanen, Fisher information and stochastic complexity, *IEEE Transactions on Information Theory* 42 (1996) 40–47.
- [25] J. Rissanen, MDL denoising, *IEEE Transactions on Information Theory* 46 (2000) 2537–2543.
- [26] J. Rissanen, *Information and Complexity in Statistical Modeling*, Springer, 2007.
- [27] T. Roos, P. Myllymäki, J. Rissanen, MDL denoising revisited, *IEEE Transactions on Signal Processing* 57 (2009) 3347–3360.
- [28] D. Schmidt, E. Makalic, MML invariant linear regression, in: A. Nicholson, X. Li (Eds.), *AI 2009: Advances in Artificial Intelligence*, 22nd Australasian Joint Conference, Melbourne, Australia, December 1–4, 2009, *Proceedings, Lecture Notes in Computer Science*, vol. 5866, Springer, 2009, pp. 312–321.
- [29] G. Schwarz, Estimating the dimension of a model, *The Annals of Statistics* 6 (1978) 461–464.
- [30] G. Seber, *A Matrix Handbook for Statisticians*, John Wiley & Sons, 2008.
- [31] G. Seber, A. Lee, *Linear Regression Analysis*, Wiley-Interscience, 2003.
- [32] A.K. Seghouane, Asymptotic bootstrap corrections of AIC for linear regression models, *Signal Processing* 90 (2010) 217–224.
- [33] A.K. Seghouane, S. Amari, The AIC criterion and symmetrizing the Kullback–Leibler divergence, *IEEE Transactions on Neural Networks* 18 (2007) 97–106.
- [34] A.K. Seghouane, M. Bekara, A small sample model selection criterion based on Kullback’s symmetric divergence, *IEEE Transactions on Signal Processing* 52 (2004) 3314–3323.
- [35] R. Shibata, Asymptotically efficient selection of the order of the model for estimating parameters of a linear process, *The Annals of Statistics* 8 (1980) 147–164.
- [36] P. Stoica, R. Moses, *Spectral Analysis of Signals*, Prentice-Hall, 2005.

- [3] Liski, A., Liski, E. P., Sund, R. and Juntunen, M. (2010)
**A comparison of WALS estimation with pretest
and model selection alternatives with an
application to costs of hip fracture treatments**
In: Yamanishi, K. et al. (eds.), *Proceedings of the Third Workshop in Information Theoretic Methods in Science and Engineering WITMSE*, August 16-18, 2010, Tampere, Finland, TICSP Series.

A COMPARISON OF WALS ESTIMATION WITH PRETEST AND MODEL SELECTION ALTERNATIVES WITH AN APPLICATION TO COSTS OF HIP FRACTURE TREATMENTS

Antti Liski¹, Erkki P. Liski², Reijo Sund³, Merja Juntunen⁴

¹ Department of Signal Processing, Tampere University of Technology,
P.O.Box 553, FIN-33101 Tampere, FINLAND, antti.liski@tut.fi

² Department of Mathematics and Statistics, University of Tampere,
FIN-33014 Tampere, FINLAND, epl@uta.fi

³ Service System Research Unit,
National Institute for Health and Welfare, FINLAND
reijo.sund@thl.fi

⁴ Centre for Health and Social Economics CHESS,
National Institute for Health and Welfare, FINLAND
merja.juntunen@thl.fi

ABSTRACT

The paper considers a new model averaging method called weighted average least squares (WALS). The method has good risk profile and its computational burden is light. The WALS technique can be easily applied to large data sets when the number of regressors is large. In the current paper the theory is used to compare the costs of hip fracture treatments between hospital districts in Finland.

1. INTRODUCTION

This paper presents a model averaging technique introduced by Magnus [9] and Magnus & Durbin [11] which is called weighted average least squares (WALS). Secondly, we apply this technique on hip fracture data of 11961 patients aged 50 or over in years 1999-2005. The purpose is to compare treatment costs of hip fracture patients between hospital districts in Finland. WALS is computationally superior over the post model selection (PMS) estimators because computing time of WALS increases only linearly with m , the number of regressors, while computing time of the PMS estimators is of order 2^m . WALS also has better risk profile over PMS estimators, and it avoids an unbounded risk. It is known that the finite-sample distributions of PMS estimators are difficult to estimate and the model selection (MS) step may have a dramatic effect on the sampling properties of PMS estimators [6].

Our framework is the ordinary linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim (0, \sigma^2 \mathbf{I}_n), \quad (1)$$

where \mathbf{X} is an $n \times p$ matrix of explanatory variables that we want to keep in the model on theoretical or other grounds. An $n \times m$ matrix \mathbf{Z} contains m additional explanatory variables which we add in the model only if they are supposed to improve estimation of $\boldsymbol{\beta}$. Following Danilov

and Magnus [3] we call the x -variables "focus" regressors and z -variables auxiliary regressors. The matrix (\mathbf{X}, \mathbf{Z}) is assumed to be of full column rank.

We have M linear regression models $\mathcal{M}_1, \dots, \mathcal{M}_M$ such that

$$\mathcal{M}_i : \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\gamma}, \sigma^2 \mathbf{I}_n\},$$

where $\mathbf{Z}_i = \mathbf{Z}\mathbf{W}_i$ and $\mathbf{W}_i = \text{diag}(w_{i1}, \dots, w_{im})$ is an $m \times m$ diagonal matrix with diagonal elements $w_{ij} \in \{0, 1\}$, $j = 1, \dots, m$. We may suppose that the models are in increasing order with respect to diagonal elements of \mathbf{W}_i when the diagonal is interpreted as m -digit binary number w_{i1}, \dots, w_{im} . Then the indices $1, \dots, M$ are associated with the diagonals as follows

$$1 \rightarrow 00 \dots 0, 2 \rightarrow 10 \dots 0, 3 \rightarrow 01 \dots 0, \\ \dots, M \rightarrow 11 \dots 1,$$

where the number of models is $M = 2^m$. For $m = 2$ we have the 4 diagonal matrices $\mathbf{W}_1 = \text{diag}(0, 0)$, $\mathbf{W}_2 = \text{diag}(1, 0)$, $\mathbf{W}_3 = \text{diag}(0, 1)$, $\mathbf{W}_4 = \text{diag}(1, 1)$ and the corresponding matrices of auxiliary regressors

$$\mathbf{Z}_i = \mathbf{Z}\mathbf{W}_i; \quad i = 1, 2, 3, 4.$$

Given an MS procedure S selecting from the set of candidate models $\mathcal{M}_1, \dots, \mathcal{M}_M$, the associated PMS estimator may be represented as

$$\hat{\boldsymbol{\beta}}_S = \sum_{i=1}^M I(S = \mathcal{M}_i) \hat{\boldsymbol{\beta}}_i, \quad (2)$$

where $\hat{\boldsymbol{\beta}}_i$ denotes the LS estimator of $\boldsymbol{\beta}$ under \mathcal{M}_i and $I(\cdot)$ is the indicator function with the value 1 for the selected model and 0 for all other models. There are

many well known MS methods such as Akaike's (AIC) and Bayesian (BIC) information criteria, as well the minimum description length (MDL) principle, for example. For (2) we have to evaluate the model goodness criterion c_S for each model \mathcal{M}_i , $1 \leq i \leq M$. If the number of variables m is large ($M = 2^m$), the task can be computationally prohibitive. By far the most common selection approach in practice is to apply a sequence of the hypothesis tests and attempt to identify the nonzero regression coefficients and select the corresponding regressors. Forward selection, backward elimination, and stepwise regression are the best known examples of these techniques [13]. All major statistical software have procedures for these techniques.

Clearly

$$\mathcal{M}_0 : \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n\} \text{ and } \mathcal{M}_M : \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}, \sigma^2 \mathbf{I}_n\},$$

where \mathcal{M}_0 is the fully restricted model without any auxiliary regressors and \mathcal{M}_M is the unrestricted model containing all auxiliary regressors. The least squares (LS) estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ under the model \mathcal{M}_M are [15] (Section 3.7)

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_0 - \mathbf{Q}\hat{\boldsymbol{\gamma}} \quad \text{and} \quad \hat{\boldsymbol{\gamma}} = (\mathbf{Z}'\mathbf{M}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{M}\mathbf{y}, \quad (3)$$

respectively, where $\hat{\boldsymbol{\beta}}_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is the LS estimator of $\boldsymbol{\beta}$ under the model \mathcal{M}_0 , $\mathbf{Q} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}$ and $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. It is known that dropping z -variables from the model decreases the variance of the LS estimator of the remaining regression parameters [7]. However, after elimination of variables, the estimates of the remaining parameters are biased if the full model is correct.

The traditional approach to select between the models \mathcal{M}_0 and \mathcal{M}_M is to test the hypothesis $\boldsymbol{\gamma} = \mathbf{0}$ and to include \mathbf{Z} if the hypothesis " $\boldsymbol{\gamma} = \mathbf{0}$ " is rejected and exclude z -variables otherwise. Then inference on $\boldsymbol{\beta}$ is made as if the resulting model were correct. In this approach the alternative estimators under consideration are the restricted LS estimator $\hat{\boldsymbol{\beta}}_0$ under the restriction $\boldsymbol{\gamma} = \mathbf{0}$ and the unbiased unrestricted LS estimator $\hat{\boldsymbol{\beta}}_M$ of $\boldsymbol{\beta}$ in the model (1). Another traditional approach is to compare the estimators $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_M$ with respect to the mean squared error (MSE) criterion. Then we test the hypothesis

$$MSE(\hat{\boldsymbol{\beta}}_0) \leq MSE(\hat{\boldsymbol{\beta}}_M) \quad (4)$$

and choose $\hat{\boldsymbol{\beta}}_0$ if the hypothesis is accepted. Here the inequality " \leq " refers to Löwner ordering of nonnegative definite matrices. Toro-Vizcarrondo and Wallace [18] made this point and developed a test for the hypothesis (4). A review of the general theory of comparing estimators under exact or stochastic linear restrictions with respect to the MSE criterion can found e.g. in Rao et.al. [14] and in Judge and Bock [5].

We shall more generally consider estimators of the model average form

$$\tilde{\boldsymbol{\beta}} = \sum_{i=1}^M c(\mathcal{M}_i) \hat{\boldsymbol{\beta}}_i, \quad (5)$$

where the weights $c(\mathcal{M}_i) \geq 0$, $1 \leq i \leq M$, sum to one and are allowed to be random, as in the post-selection estimator class. Buckland [1] suggested using weights proportional to $\exp(-AIC_i/2)$, where AIC_i is the AIC score for the candidate model \mathcal{M}_i . Similar weighting can be derived from other model selection criteria as well. Liski et al. [8] proposed a model average estimator using weights derived from the MDL criterion. However, for large values of m the estimator (5) is infeasible unless the set of candidate models is somehow restricted.

2. PRETESTING

For simplicity we assume for a moment that $m = 1$, and consequently \mathbf{Z} is a single $n \times 1$ vector z . Then we have two alternative models $\mathcal{M}_0, \mathcal{M}_1$, where \mathcal{M}_0 is the restricted model as before and

$$\mathcal{M}_1 : \mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta} + z\boldsymbol{\gamma}, \sigma^2 \mathbf{I})$$

is the unrestricted model. At first we assume that σ^2 is known, but later in WALS implementation (Section 4) σ^2 is replaced by its usual unbiased estimator obtained from the unrestricted model. Using the notation

$$\mathbf{q} = \frac{\sigma}{\sqrt{z'\mathbf{M}z}} (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'z \quad \text{and} \quad \theta = \frac{\boldsymbol{\gamma}}{\sigma/\sqrt{z'\mathbf{M}z}},$$

we can write the LS estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ under \mathcal{M}_1 as

$$\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_0 - \hat{\theta}\mathbf{q}, \quad \hat{\boldsymbol{\gamma}} = \frac{z'\mathbf{M}\mathbf{y}}{z'\mathbf{M}z},$$

where $\hat{\theta} = \frac{\hat{\boldsymbol{\gamma}}}{\sigma/\sqrt{z'\mathbf{M}z}}$ denotes the t -ratio, which follows the normal distribution $N(\theta, 1)$ because σ^2 is known. Note that $\hat{\theta}$ and $\hat{\boldsymbol{\beta}}_0$ are independent.

The traditional approach to choose between $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_1$ is to use the t -ratio. If $\hat{\theta}$ is large, we choose the unrestricted LS estimator $\hat{\boldsymbol{\beta}}_1$, and if $\hat{\theta}$ is small, we choose the restricted LS estimator $\hat{\boldsymbol{\beta}}_0$. This leads to the estimator

$$\tilde{\boldsymbol{\beta}}_{pre} = \begin{cases} \hat{\boldsymbol{\beta}}_0 & \text{if } |\hat{\theta}| \leq c; \\ \hat{\boldsymbol{\beta}}_1 & \text{if } |\hat{\theta}| > c, \end{cases}$$

where c is some positive constant. For example, $c = 1.96$ corresponds to the 5% significance level.

Given that we are interested in the best possible estimation on $\boldsymbol{\beta}$, not in $\boldsymbol{\gamma}$, the proper question of interest is, "Is $\hat{\boldsymbol{\beta}}_0$ better estimator than $\hat{\boldsymbol{\beta}}_1$?" When assessing estimators with respect to their MSE , we should compare their MSE matrices, which are

$$MSE(\hat{\boldsymbol{\beta}}_0) = \text{Cov}(\hat{\boldsymbol{\beta}}_0) + \theta^2 \mathbf{q}\mathbf{q}' \quad \text{and}$$

$$MSE(\hat{\boldsymbol{\beta}}_1) = \text{Cov}(\hat{\boldsymbol{\beta}}_1) + \text{Var}(\hat{\theta}) \mathbf{q}\mathbf{q}' = \text{Cov}(\hat{\boldsymbol{\beta}}_0) + \mathbf{q}\mathbf{q}'.$$

Then

$$MSE(\hat{\boldsymbol{\beta}}_0) - MSE(\hat{\boldsymbol{\beta}}_1) = (\theta^2 - 1) \mathbf{q}\mathbf{q}',$$

where \mathbf{q} is a known vector and θ is the usual non-centrality parameter associated with the t -ratio for testing $\boldsymbol{\gamma} = \mathbf{0}$.

Hence (cf. [9])

$$\begin{aligned} MSE(\hat{\beta}_0) &\leq MSE(\hat{\beta}_1) & \text{if } \theta^2 < 1 \\ MSE(\hat{\beta}_0) &= MSE(\hat{\beta}_1) & \text{if } \theta^2 = 1 \\ MSE(\hat{\beta}_0) &\geq MSE(\hat{\beta}_1) & \text{if } \theta^2 > 1. \end{aligned}$$

Toro-Vizcarrondo and Wallace [18] obtained a uniformly most powerful test for $H_0 : \theta^2 \leq 1$ vs. $H_1 : \theta^2 > 1$ from the probability

$$P(|\hat{\theta}| \leq c | \theta^2 = 1) = 1 - \alpha,$$

where α denotes the significance level. E.g. a 5% test corresponds to $c = 2.65$.

There are two problems in applying either the usual t-test or the test of Toro-Vizcarrondo and Wallace. The first is that the choice of significance is largely arbitrary. In the preliminary test we decide whether to use $\hat{\beta}_0$ or $\hat{\beta}_1$. The second problem is that after the test neither $\hat{\beta}_0$ nor $\hat{\beta}_1$ is actually used. The estimator actually used is the pretest estimator

$$\tilde{\beta}_{pre} = \lambda \hat{\beta}_1 + (1 - \lambda) \hat{\beta}_0, \quad (6)$$

where $\lambda = 1$ if $|\hat{\theta}| > c$, and otherwise $\lambda = 0$.

More generally, any estimator of the form

$$\tilde{\beta} = \lambda \hat{\beta}_1 + (1 - \lambda) \hat{\beta}_0, \quad 0 \leq \lambda \leq 1 \quad \text{and } \lambda = \lambda(\hat{\theta}) \quad (7)$$

will be called a weighted average least squares (WALS) estimator [11]. The pretest estimator is an example of such an estimator. Usually λ is a nondecreasing function of $|\hat{\theta}|$, so that the larger $|\hat{\theta}|$ the larger λ will be and hence more weight will be put on $\hat{\beta}_1$ relative to $\hat{\beta}_0$. It turns out that the pretest estimators (6) have poor properties [9] and better estimators can be found in the wider class of WALS estimators [10].

The following equivalence theorem, originally proved by Magnus and Durbin [11] and extended by Danilov and Magnus [3] turns out useful in the study of WALS estimators.

Theorem 1. (Equivalence theorem) Let $\tilde{\beta} = \lambda \hat{\beta}_1 + (1 - \lambda) \hat{\beta}_0$ be a WALS estimator of β and $\tilde{\theta} = \lambda \hat{\theta}$, where λ is as in (7). Then

$$\begin{aligned} E(\tilde{\beta}) &= \beta - E(\tilde{\theta} - \theta)q, \\ \text{Var}(\tilde{\beta}) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \text{Var}(\tilde{\theta})qq' \text{ and hence} \\ MSE(\tilde{\beta}) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + MSE(\tilde{\theta})qq'. \end{aligned}$$

The equivalence theorem expresses the expectation, variance and mean square error of a WALS estimator $\tilde{\beta}$ of β as a function of the corresponding characteristics of the estimator $\tilde{\theta}$ of θ . Thus $\text{Var}(\tilde{\beta})$ and $MSE(\tilde{\beta})$ are minimized if and only if $\text{Var}(\tilde{\theta})$ and $MSE(\tilde{\theta})$, respectively, are minimized.

Theorem 1 is important because it shows that studying the WALS estimator (7) for the regression problem is equivalent to studying the estimator $\tilde{\theta} = \lambda(\hat{\theta})\hat{\theta}$ of θ in the simple normal distribution $N(\theta, 1)$. Suppose that

we have found an optimal estimator $\tilde{\theta}_0 = \lambda_0(\hat{\theta})\hat{\theta}$ of θ . Then the equivalence theorem guarantees that this same λ -function λ_0 will provide the optimal WALS estimator (7) of β . Thus the problem is to find an optimal λ -function. When the risk of an estimator $\tilde{\theta}$ is defined as its MSE,

$$R(\theta; \tilde{\theta}) = E_\theta(\tilde{\theta} - \theta)^2, \quad (8)$$

Magnus [9] showed that the traditional pretest estimator (6) has many undesirable risk properties and the WALS estimators (7) have advantages (cf. [10]) over the pretest estimators (6).

3. THE LAPLACE ESTIMATOR

Let $y \sim N(\theta, 1)$ and let $T(y) = \lambda y$ be an estimator of θ , where $\lambda = \lambda(y)$ is a scalar function of y such that $0 \leq \lambda(y) \leq 1$. The ordinary LS estimator (and the ML estimator) of θ is $T(y) = y$ with $\lambda \equiv 1$ whereas the traditional pretest estimator is obtained when $\lambda(y) = 1$ if $|y| > c$ and $\lambda(y) = 0$ if $|y| \leq c$ for some fixed threshold value $c > 0$. Note that $T(y) = \lambda y$ is a weighted average of the LS estimator y (corresponds to $\hat{\beta}_1$) and the null estimator $T(y) \equiv 0$ (corresponds to $\hat{\beta}_0$).

Our aim is to find a good WALS estimator of the regression parameter β with respect to the MSE criterion. The equivalence theorem tells us that a good estimator $T(y) = \lambda y$ of θ under the model $N(\theta, 1)$ will allow us to make a good estimator of β . Typically $0 \leq \lambda \leq 1$ is a nondecreasing function of $|y|$ so that $T(y) = \lambda y$ shrinks the LS estimator y towards zero. Magnus [10] considered a wide range of possible estimators of θ and finally he preferred the Laplace estimator which is admissible, has bounded risk, has good properties around $|\theta| = 1$, and is near optimal in terms of minimax regret. The value $|\theta| = 1$ is an important pivot since the risk of the null estimator of θ is less than the risk of the LS estimator y if and only if $|\theta| < 1$ ([9], [10]).

Assuming a Laplace prior density $\frac{1}{2}c \exp(-c|\theta|)$, $-\infty < \theta < \infty$, the posterior mean and variance of θ given y can be written as [12]

$$\begin{aligned} E(\theta|y) &= \frac{1 + h(y)}{2}(y - c) + \frac{1 - h(y)}{2}(y + c) \\ &= y - h(y)c, \end{aligned} \quad (9)$$

$$\text{Var}(\theta|y) = 1 + c^2[1 - h^2(y)] - \frac{c[1 + h(y)]\phi(y - c)}{\Phi(y - c)},$$

where

$$h(y) = \frac{1 - e^{2cy}\Psi(y)}{1 + e^{2cy}\Psi(y)}, \quad \Psi(y) = \frac{\Phi(-y - c)}{\Phi(y - c)}$$

and $\phi(\cdot)$ denotes the density and $\Phi(\cdot)$ the distribution function of the standard normal distribution, respectively. The hyperparameter c is chosen $c = \log 2$ which implies that $\text{median}(\theta) = 0$ and $\text{median}(\theta^2) = 1$. The posterior mean (9) is the Laplace estimator

$$L(y) = y - h(y)c = \lambda(y)y \quad (10)$$

with $\lambda(y) = (1 - \frac{h(y)c}{y})$. The function $h(\cdot)$ is monotonically increasing with $h(-\infty) = -1$, $h(0) = 0$, $h(\infty) = 1$ and $h(-y) = -h(y)$, and hence $\lambda(y) = \lambda(-y) \rightarrow 1$ as $y \rightarrow \infty$. It can be shown that $h(y) \rightarrow 0.58956$ as $y \rightarrow 0$. Magnus [10] and Danilov [2] have studied the properties of the Laplace estimator in detail.

4. WALS ESTIMATION

4.1. Restricted LS estimators

We can always find an orthogonal $m \times m$ matrix \mathbf{P} such that $\mathbf{P}'\mathbf{Z}'\mathbf{M}\mathbf{Z}\mathbf{P} = \mathbf{\Lambda}$ is diagonal and define new auxiliary regressors $\mathbf{Z}^* = \mathbf{Z}\mathbf{P}\mathbf{\Lambda}^{-1/2}$ and new auxiliary parameters $\gamma^* = \mathbf{\Lambda}^{1/2}\mathbf{P}'\gamma$ as noted by Magnus et al. [12]. Hence there is no loss of generality to posit

$$\mathbf{Z}'\mathbf{M}\mathbf{Z} = \mathbf{I}_m. \quad (11)$$

We assume in the sequel that (11) holds. Then it follows from (3) that

$$\hat{\gamma} = \mathbf{Z}'\mathbf{M}\mathbf{y} \quad \text{and} \quad \hat{\gamma} \sim N(\gamma, \sigma^2\mathbf{I}_m).$$

In general, given the assumption (11), the restricted LS estimator for β under the model \mathcal{M}_i , $1 \leq i \leq M$, is

$$\hat{\beta}_i = \hat{\beta}_0 - \mathbf{Q}\mathbf{W}_i\hat{\gamma}. \quad (12)$$

4.2. The WALS estimator of β

The WALS estimator of β is defined as

$$\tilde{\beta} = \sum_{i=1}^M v_i \hat{\beta}_i, \quad (13)$$

where the weight functions v_i satisfy the conditions

$$v_i \geq 0, \quad \sum_i v_i = 1, \quad v_i = v_i(\hat{\gamma}). \quad (14)$$

It follows from (12) and (13) that the WALS estimator can be written as

$$\tilde{\beta} = \hat{\beta}_0 - \mathbf{Q}\mathbf{W}\hat{\gamma}, \quad (15)$$

where $\mathbf{W} = \sum_i v_i \mathbf{W}_i$ is a diagonal random matrix such that

$$\hat{\gamma}'\mathbf{W} = (\lambda_1\hat{\gamma}_1, \dots, \lambda_m\hat{\gamma}_m) \quad \text{with} \quad \hat{\gamma}_j \sim N(\gamma_j, \sigma^2).$$

We choose $\lambda_j = \lambda_j(\hat{\gamma}_j)$, $1 \leq j \leq m$. Since $\hat{\gamma}_1, \dots, \hat{\gamma}_m$ are independent, it follows that $\lambda_1, \dots, \lambda_m$ are independent. Hence we have m identical one-dimensional problems to estimate the elements λ_j .

4.3. LAPLACE weights

If we denote $\theta = \gamma/\sigma$ and compute $\hat{\theta}$, then the elements $\hat{\theta}_1, \dots, \hat{\theta}_m$ of $\hat{\theta}$ are independent and $\hat{\theta}_j \sim N(\theta_j, 1)$, $1 \leq j \leq m$. Then we have identical estimation problems in the models $N(\theta_j, 1)$. By (10) and (9) we can compute the Laplace estimates $\tilde{\theta}_j = L(\hat{\theta}_j)$ and their variances $\phi_j^2 = \text{Var}(\theta_j|\hat{\theta}_j)$, $1 \leq j \leq m$. We define $\tilde{\theta} =$

$(\tilde{\theta}_1, \dots, \tilde{\theta}_m)$, $\Phi = \text{diag}(\phi_1^2, \dots, \phi_m^2)$ and note that $\gamma = \sigma\theta$. Then the WALS estimators for β and γ with Laplace weights can be computed as

$$\tilde{\gamma} = \sigma\tilde{\theta} \quad \text{and} \quad \tilde{\beta} = \hat{\beta}_0 - \mathbf{Q}\tilde{\gamma}.$$

The variance of $\tilde{\gamma}$ and $\tilde{\beta}$ is

$$\begin{aligned} \text{Var}(\tilde{\gamma}) &= \sigma^2\Phi \quad \text{and} \\ \text{Var}(\tilde{\beta}) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{Q}\text{Var}(\tilde{\gamma})\mathbf{Q}'. \end{aligned}$$

The results above are based on the assumption that σ^2 is known, but in practice σ^2 is unknown and it must be estimated from data. This problem is solved by replacing σ^2 by its usual unbiased s^2 obtained from the unrestricted model. Danilov [2] showed that this approximation is very accurate and the main properties of the Laplace estimator change only marginally.

4.4. The equivalence theorem

The equivalence theorem proved by Danilov and Magnus ([3], Theorem 1) states that if the assumption (11) holds and the conditions (14) on weight functions are satisfied, then

$$\begin{aligned} \text{Var}(\tilde{\beta}) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{Q}\text{Var}(\mathbf{W}\hat{\gamma})\mathbf{Q}' \\ \text{MSE}(\tilde{\beta}) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{Q}\text{MSE}(\mathbf{W}\hat{\gamma})\mathbf{Q}'. \end{aligned}$$

Now $\text{Var}(\tilde{\beta})$ depends only on $\text{Var}(\mathbf{W}\hat{\gamma})$ and $\text{MSE}(\tilde{\beta})$ only on $\text{MSE}(\mathbf{W}\hat{\gamma})$.

5. MEDICAL CARE COSTS OF HIP FRACTURE TREATMENTS

Hip fracture is a common and important cause leading to lowered mobility or ultimately to death among the elderly population. In Finland, the number of hip fractures in people aged 50 or over was on average 5564 between the years 1998-2002 ([17]). Not only patients suffer from hip fractures, but they also cause remarkable costs to society. The costs of treating a hip fracture patient are about three-fold compared to the caring for a patient without a fracture ([4]).

Comparison of treatments and outcomes between medical centres treating hip fractures can yield information for the development of treatment and serve as a quality assessment of care. Profiling medical care providers on the basis of quality of care and utilization of resources has become a widely used analysis in health care policy and research. Risk-adjustment is desirable when comparing hospitals or hospital districts with respect to a performance indicator such as treatment cost. Adjustment is intended to account for possible differences in patient case mix

This paper presents a model for hip fracture treatment costs in Finland using linear regression. Data were obtained by combining from several national registries [16] and consisted of 36492 patients aged 50 or over from the years 1999-2005. There are 21 hospital districts in Finland but here we report only results of the seven largest hospital districts. We concentrate only on patients who have

not been institutionalised before the fracture and were not institutionalised after the fracture but were able to return home after the treatment. Patients who died within a year after the fracture were removed from the data. After all the exclusions, the data set used in this paper contained 11961 patients.

The dependent variable in the model is the treatment cost. The model contains seven focus regressors and 38 auxiliary regressors. The seven largest hospital districts are chosen as the focus regressors because we wish to test whether there is difference in treatment costs between the hospital districts. The hospital district of Helsinki and Uusimaa is taken as baseline. The set of auxiliary regressors contains a number of important comorbidities like congestive heart failure, diabetes and cancer. The auxiliary regressors are intended to reflect the mix of patients treated by a hospital or unit. The focus regressors are given in Table 1 and the auxiliary regressors are explained in Table 3.

6. ESTIMATION RESULTS

We estimate the model using WALs as discussed in Section 4. We also estimated the model applying a backward elimination (BE) technique. In BE we start using the full model with all $p + m$ variables and we eliminate auxiliary variables having the smallest F statistic but we never add regressors (Matlab's stepwise fit routine). The selected model thus contains all focus regressors and a subset of the auxiliary regressors. The reported LS estimates and standard errors are thus conditional on the model selected. BE selection procedure is considered because stepwise selection procedures are commonly used in practice.

The guidelines for hip fracture treatment are the same in the whole country and therefore we assume that all hospital districts treat the patients according to the same standards. In Table 1 all four estimation techniques indicate differences in treatment costs between hospital districts. The results of WALs seem to be in agreement with those of BE and the unrestricted LS whereas the restricted LS estimates show some discrepancy. The restricted LS indicates that costs in Central Finland are significantly higher than in Helsinki and Uusimaa but the other methods do not support the difference. According to the restricted LS the costs in Central Finland are significantly higher than than in Helsinki and Uusimaa but the other methods do not indicate such a difference. Satakunta hospital district have the highest and Northern Savo the lowest treatment costs, and these costs differ significantly from the costs in Helsinki and Uusimaa. By cost the most significant auxiliary variables are age, waiting for operation over 2 days, Parkinson disease, alcohol abuse, hypertension and diabetes.

The limited practical experience in the use of WALs seems to show that BE has a tendency to give larger absolute t -values on the average than WALs [12]. This tendency is mildly visible also in our results (Table 1 and Table 2). The variances of the BE estimates are conditional on the set of variables selected and the conditional

Table 1. Estimates $\hat{\beta}$, standard errors (in parentheses) and t -values of the focus regressors.

Variable	$\hat{\beta}_W$	t_W	$\hat{\beta}_{BE}$	t_{BE}
Helsinki and Uusimaa	2331.04(307.08)	7.59	2004.82(307.55)	6.52
SW Finland	-65.70(117.75)	-0.56	-78.26(117.59)	-0.67
Satakunta	484.97(147.07)	3.30	451.30(146.90)	3.07
Pirkanmaa	-128.44(115.09)	-1.12	-148.18(114.93)	-1.29
N Savo	-826.07(143.74)	-5.75	-841.63(143.67)	-5.86
C Finland	142.92(142.79)	1.00	118.55(142.64)	0.83
N Ostrobothnia	-348.58(136.21)	-2.56	-373.42(136.26)	-2.74
Variable	Unrestricted	t_u	Restricted	t_r
Helsinki and Uusimaa	1927.96(310.16)	6.22	9052.56(63.50)	142.56
SW Finland	-69.33(117.82)	-0.59	45.24(122.16)	0.37
Satakunta	454.98(147.19)	3.09	825.78(151.24)	5.46
Pirkanmaa	-140.60(115.16)	-1.22	118.31(119.09)	0.99
N Savo	-831.68(143.80)	-5.78	-662.68(149.20)	-4.44
C Finland	128.79(142.90)	0.90	395.95(148.04)	2.67
N Ostrobothnia	-368.96(136.42)	-2.70	21.40(139.62)	0.15

SW-Southwest, N-Northern, C-Central

Table 2. Estimates and t -values of the auxiliary regressors using WALs, the unrestricted LS and BE.

Regressor	$\hat{\beta}_W$	t_W	$\hat{\beta}_M$	t_M	$\hat{\beta}_{BE}$	t_{BE}
HOSP90	8.35	2.16	7.20	1.82	7.93	2.01
AGE	81.41	19.82	86.02	20.77	85.79	20.87
HOSPDUM	284.53	2.77	351.88	3.17	380.24	3.46
OPWAIT	1570.97	12.73	1688.85	13.55	1689.15	13.57
FEMALE	-230.18	-2.60	-261.29	-2.90	-274.75	-3.09
CHF	313.91	3.00	280.90	2.41	318.79	2.77
Arr	188.00	1.78	267.79	2.33	288.89	2.54
Val	73.11	0.36	105.69	0.41	0	0
PCD	191.37	0.65	324.99	0.89	0	0
PVD	223.16	1.28	222.19	1.19	0	0
Par	922.27	2.18	1231.32	2.85	1251.85	2.90
PaD	1078.49	5.41	1198.46	5.56	1186.85	5.52
Dem	333.46	2.32	384.28	2.49	377.36	2.45
OND	490.59	2.74	617.27	3.19	630.53	3.27
CPD	193.30	1.74	229.24	1.88	275.30	2.29
Hyp	112.95	0.86	121.86	0.73	0	0
Ren	513.71	1.43	745.64	1.93	798.21	2.08
LID	95.38	0.26	168.56	0.36	0	0
PUD	190.99	0.65	307.69	0.88	0	0
Lym	82.58	0.17	41.17	0.06	0	0
McA	171.16	0.33	184.85	0.28	0	0
STu	-28.76	-0.21	-9.20	-0.06	0	0
Rhe	313.92	2.82	326.22	2.39	352.58	2.59
Coa	613.41	0.71	996.15	0.95	0	0
Obe	1253.28	1.01	1948.63	1.35	0	0
WeL	61.93	0.12	32.80	0.05	0	0
FED	354.99	1.83	496.26	2.09	521.40	2.21
BLA	1846.60	3.24	2270.21	3.98	2281.51	4.03
DeA	-61.26	-0.30	-21.97	-0.09	0	0
Alc	1023.23	5.57	1218.61	5.78	1297.83	6.27
Dru	568.81	0.98	897.39	1.33	0	0
Psy	619.72	3.90	730.95	4.23	747.53	4.39
Dep	34.08	0.21	56.95	0.32	0	0
Pne	178.14	1.40	230.52	1.69	0	0
UTI	-110.77	-0.78	-145.45	-0.90	0	0
Inj	142.79	1.63	110.72	1.25	0	0
Hyt	369.91	4.58	396.62	4.87	405.76	5.00
Dia	407.89	3.74	393.08	3.26	407.41	3.40

estimates may be spuriously precise resulting in misleadingly high t -values. In our data the auxiliary variables are not correlated or only weakly, therefore WALs, BE and the unrestricted LS does not differ much.

We emphasize that even though we have found statistically significant differences in the treatment costs between the hospital districts, inferences concerning these results should be made with caution. Before drawing conclusions on the performance of the hospital districts a more careful and extensive study and interpretation of the results should be conducted with experts from the medical field participating actively in the research. These methods will be applied in this wider study and the work is in progress. The present application is a pilot study.

Table 3. Explanation of the auxiliary regressors.

Variable	Explanation	Variable	Explanation
HOSP90	days spent in hospital before the fracture	Lym	Lymphoma*
AGE	age of the patient	MCa	Metastatic cancer*
HOSPDUM	hospitalized during 90 days before the fracture*	STu	Solid tumor without metastasis*
OPWAIT	waited for operation over 2 days*	Rhe	Rheumatoid arthritis*
FEMALE	patient is a female*	Coa	Coagulopathy*
CHF	Congestive heart failure*	Obe	Obesity*
Arr	Cardiac arrhythmias*	WeL	Weight loss*
Val	Valvular disease*	FED	Fluid and electrolyte disorders*
PCD	Pulmonary circulation disorders*	BLA	Blood loss anemia*
PVD	Peripheral vascular disorders*	DeA	Deficiency anemia*
Par	Paralysis*	Alc	Alcohol abuse*
PaD	Parkinson disease*	Dru	Drug abuse*
Dem	Dementia*	Psy	Psychoses*
OND	Other neurological disorders*	Dep	Depression*
CPD	Chronic pulmonary disorders*	Pne	Pneumonia*
Hyp	Hypothyroidism*	UTI	Urinary tract infection*
Ren	Renal failure*	Inj	Injuries*
LiD	Liver disease*	Hyt	Hypertension*
PUD	Peptic ulcer disease*	Dib	Diabetes*

dummy variables are marked with *

7. ACKNOWLEDGMENTS

The work of Reijo Sund was financially supported by the Yrjö Jahnsson Foundation (grant 5978).

8. REFERENCES

- [1] Buckland, S. T. Burnham, K. P. and Augustin, N. H. (1999). Model Selection: An Integral Part of Inference. *Biometrics*, 53, 603–618.
- [2] Danilov, D. (2005). Estimation of the mean of a univariate normal distribution when the variance is not known. *Econometrics Journal* 8, 277–291.
- [3] Danilov, D. and Magnus, J. R. (2004). On the harm that ignoring pretesting can cause. *Journal of Econometrics*, 122, 27–46.
- [4] Haentjens, P., Autier, P., Barette, M., Boonen, S. and Belgian Hip Fracture Study Group (2001) The economic cost of hip fractures among elderly women. A one-year, prospective, observational cohort study with matched-pair analysis. *J Bone Joint Surg Am*, 83-A, 493–500.
- [5] Judge, G. G., and Bock, M. E. Bock *The statistical implications of pre-test and Stein-rule estimators in econometrics*, Amsterdam, North-Holland.
- [6] Leeb, H. and Pötker, B. M. (2005) Model selection and inference: facts and fiction. *Econometric Theory*, 21, 21–59.
- [7] Liski, E. P. and Trenkler, G. (1993). MSE-Improvement of the Least Squares Estimator by Dropping Variables. *Metrika* 40, 263–269.
- [8] Liski, E. P. and Liski, A. (2008). MDL Model Averaging for Linear Regression. In: Grüwald, P., Myllymäki, P., Tabus, I., Weinberger, M., and Yu, B. (Eds.). *Festschrift in Honor of Jorma Rissanen on the Occasion of his 75th Birthday*, 145–154. Tampere, Tampere International Center for Signal Processing.
- [9] Magnus, J. R. (1999). The traditional pretest estimator. *Theory of Probability and Its Applications*, 44, 293–308.
- [10] Magnus, J. R. (2002). Estimation of the mean of a univariate normal distribution with a known variance. *Econometrics Journal*, 5, 225–236.
- [11] Magnus, J. R. and Durbin, J. (1999). Estimation of regression coefficients of interest when other regression coefficients are of no interest. *Econometrica*, 67, 639–643.
- [12] Magnus, J. R., Powell, O. and Prüfer, P. (2010). A comparison of two model averaging techniques with an application to growth empirics. *Journal of Econometrics*, 154, 139–153.
- [13] Miller, A. (2002) *Subset selection in regression*, 2nd ed. London, Chapman & Hall/CRC.
- [14] Rao, C. R., Toutenburg, H., Shalabh and Heuman, C. (2008). *Linear Models and Generalizations. Least squares and Alternatives*, 3rd ed. Berlin, Springer.
- [15] Seber, G. A. F. and Lee, A. J. (2003). *Linear Regression Analysis*, 2nd ed. New York, Wiley.
- [16] Sund, R., Juntunen, M., Lüthje, P., Huusko, T., Mäkelä, M., Linna, M., Liski, A., Häkkinen, U. (2006). *PERFECT - Hip Fracture, Performance, Effectiveness and Cost of Hip Fracture Treatment Episodes* (In Finnish), National Research and Development Centre for Welfare and Health, Helsinki.
- [17] Sund, R. (2007) Utilization of routinely collected administrative data in monitoring the incidence of aging dependent hip fracture. *Epidemiologic Perspectives & Innovations*, 2007, 4:2. <http://www.epi-perspectives.com/content/4/1/2>
- [18] Toro-Vizcarrondo, C. and Wallace, W. D. (1968). A test of the mean square error criterion for restrictions in linear regression. *Journal of the American Statistical Association*, 63, pp. 558–572.

- [4] Liski, A., Liski, E. P. and Häkkinen, U. (2013)
**Shrinkage estimation via penalized least squares
in linear regression with an application to hip
fracture treatment costs**
Proceedings of the 9th Tartu Conference on Multivariate Statistics, World
Scientific, Accepted for publication.

Shrinkage estimation via penalized least squares in linear regression with an application to hip fracture treatment costs

ANTTI LISKI

*Department of Signal Processing, Tampere University of Technology,
Tampere, Finland*

**E-mail: antti.liski@tut.fi*

ERKKI P. LISKI

*Department of Mathematics and Statistics, University of Tampere,
Tampere, Finland*

E-mail: erkki.liski@uta.fi

UNTO HÄKKINEN

*Centre for Health and Social Economics, National Institute for Health and Welfare,
Helsinki, Finland*

E-mail: unto.hakkinen@thl.fi

In this paper, we consider the problem of averaging across least squares estimates obtained from a set of models. Existing model averaging (MA) methods usually require estimation of a single weight for each candidate model. However, in applications the number of candidate models may be huge. Then the approach based on estimation of single weights becomes computationally infeasible. Utilizing a connection between shrinkage estimation and model weighting we present an accurate and computationally efficient MA estimation method. The performance of our estimators is displayed in simulation experiments which utilize a realistic set up based on real data.

Keywords: Model averaging, Model selection, Mean square error, Efficiency bound, Simulation experiment

1. Introduction

Our framework is the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 \mathbf{I}_n), \quad (1)$$

where \mathbf{X} and \mathbf{Z} are $n \times p$ and $n \times m$ matrices of nonrandom regressors, (\mathbf{X}, \mathbf{Z}) is assumed to be of full column-rank $p + m < n$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are $p \times 1$

and $m \times 1$ vectors of unknown parameters. Our interest is in the effect of \mathbf{X} on \mathbf{y} , that is, we want to estimate $\boldsymbol{\beta}$ while the role of \mathbf{Z} is to improve the estimation of $\boldsymbol{\beta}$. It is known that dropping z -variables from the model decreases the variance of the least squares (LS) estimator of the β -parameters. However, after elimination of variables, the estimates are biased if the full model is correct. In certain applications the model (1) can also be interpreted as an analysis of covariance (ANCOVA) model which is a technique that sits between the analysis of variance and regression analysis. However, ANCOVA is only a special instance of the general regression model (1).

We introduce a set of shrinkage estimators for the regression coefficients $\boldsymbol{\beta}$ in the class of penalized least squares estimators. The efficiency bound of estimators with respect to the mean square (*MSE*) error criterion within this shrinkage class is known. We search for the estimators whose *MSE* is uniformly close to the efficiency bound. It turns out that many interesting known estimators belong to this class, for example the soft thresholding and the firm thresholding estimators, non-negative garrote, LASSO and SCAD estimators. On the other hand, for example the hard thresholding rule (pre testing) and the ridge estimator do not belong to this shrinkage class. In Section 2 we present the canonical form of the model (1). The problem of model selection and averaging is introduced in Section 3. We characterize our class of shrinkage estimators within the set of penalized least squares estimators in Subsection 4.1, and the main result on shrinkage and penalised least squared estimation is given in Subsection 4.2. Examples of good alternative penalized least squares estimators, which are also shrinkage estimators, are introduced in Subsection 4.3. A real data application is given in Section 5 and the results of the simulation experiments are reported in Section 6.

2. The model

We will work with the canonical form of the model (1) where z -variables are orthogonalized by writing the systematic part of the model (1) as

$$\begin{aligned}\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{P}\mathbf{Z}\boldsymbol{\gamma} + (\mathbf{I} - \mathbf{P})\mathbf{Z}\boldsymbol{\gamma} \\ &= \mathbf{X}\boldsymbol{\alpha} + \mathbf{M}\mathbf{Z}\boldsymbol{\gamma},\end{aligned}\tag{2}$$

where

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \quad \text{and} \quad \mathbf{M} = \mathbf{I}_n - \mathbf{P}\tag{3}$$

are symmetric idempotent matrices and $\boldsymbol{\alpha} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\boldsymbol{\gamma}$. Since $(\mathbf{M}\mathbf{Z})'\mathbf{M}\mathbf{Z} = \mathbf{Z}'\mathbf{M}\mathbf{Z}$ is positive definite,¹ then there exists a nonsingular

matrix \mathbf{C} such that²

$$\mathbf{C}'\mathbf{Z}'\mathbf{M}\mathbf{Z}\mathbf{C} = (\mathbf{M}\mathbf{Z}\mathbf{C})'(\mathbf{M}\mathbf{Z}\mathbf{C}) = \mathbf{U}'\mathbf{U} = \mathbf{I}_m. \quad (4)$$

In (4) $\mathbf{U} = \mathbf{M}\mathbf{Z}\mathbf{C}$ denotes the matrix of orthogonal canonical auxiliary regressors. Introducing the canonical auxiliary parameters $\boldsymbol{\theta} = \mathbf{C}^{-1}\boldsymbol{\gamma}$ we can write in (2)

$$\mathbf{M}\mathbf{Z}\boldsymbol{\gamma} = \mathbf{M}\mathbf{Z}\mathbf{C}\mathbf{C}^{-1}\boldsymbol{\gamma} = \mathbf{U}\boldsymbol{\theta}.$$

There are advantages working with $\boldsymbol{\theta}$ instead of $\boldsymbol{\gamma}$, and we can always get $\boldsymbol{\gamma}$ from $\boldsymbol{\gamma} = \mathbf{C}\boldsymbol{\theta}$.

We are interested in estimation of $\boldsymbol{\beta}$, whereas $\boldsymbol{\theta}$ contains auxiliary parameters. Let \mathcal{M}_0 denote the fully restricted model without any auxiliary regressors and \mathcal{M}_M the unrestricted model containing all auxiliary regressors as follows

$$\mathcal{M}_0 : \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\} \text{ and } \mathcal{M}_M : \{\mathbf{y}, \mathbf{X}\boldsymbol{\alpha} + \mathbf{U}\boldsymbol{\theta}, \sigma^2\mathbf{I}_n\}. \quad (5)$$

Now in θ -parametrization we write $\boldsymbol{\alpha} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{C}\mathbf{C}^{-1}\boldsymbol{\gamma} = \boldsymbol{\beta} + \mathbf{Q}\boldsymbol{\theta}$, where $\mathbf{Q} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{C}$. The model \mathcal{M}_M is orthogonal such that $\mathbf{X}'\mathbf{U} = \mathbf{0}$ and (\mathbf{X}, \mathbf{U}) is of full column rank. Then the least squares (LS) estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ from the model \mathcal{M}_M are

$$\begin{aligned} \hat{\boldsymbol{\alpha}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \\ \hat{\boldsymbol{\theta}} &= \mathbf{U}'\mathbf{y}. \end{aligned}$$

Let $\hat{\boldsymbol{\beta}}_0$ denote the LS estimator of $\boldsymbol{\beta}$ under the restricted model \mathcal{M}_0 and note that $\hat{\boldsymbol{\alpha}} \equiv \hat{\boldsymbol{\beta}}_0$. The correspondence between the vectors $(\boldsymbol{\alpha}', \boldsymbol{\theta}')$ and $(\boldsymbol{\beta}', \boldsymbol{\theta}')$ is one-to-one, and consequently the same correspondence holds between their LS estimates. Hence the LS estimate of $\boldsymbol{\beta}$ under the unrestricted model \mathcal{M}_M is^{1,3}

$$\begin{aligned} \hat{\boldsymbol{\beta}}_M &= \hat{\boldsymbol{\alpha}} - \mathbf{Q}\hat{\boldsymbol{\theta}} \\ &= \hat{\boldsymbol{\beta}}_0 - \mathbf{Q}\hat{\boldsymbol{\theta}}. \end{aligned}$$

In the unrestricted model \mathcal{M}_M in (5) there are m components of $\boldsymbol{\theta}$, and 2^m submodels are obtained by setting various subsets of the elements $\theta_1, \dots, \theta_m$ of $\boldsymbol{\theta}$ equal to zero. These 2^m models $\mathcal{M}_0, \dots, \mathcal{M}_M$ can be written as

$$\mathcal{M}_i : \{\mathbf{y}, \mathbf{X}\boldsymbol{\alpha} + \mathbf{U}_i\boldsymbol{\theta}, \sigma^2\mathbf{I}_n\},$$

where $\mathbf{U}_i = \mathbf{U}\mathbf{W}_i$ and $\mathbf{W}_i = \text{diag}(w_{i1}, \dots, w_{im})$, $i = 0, 1, \dots, M$ are $m \times m$ diagonal matrices with diagonal elements $w_{ij} \in \{0, 1\}$, $j = 1, \dots, m$, and $M = 2^m - 1$. We may suppose that the models are in increasing order with

respect to diagonal elements of \mathbf{W}_i when the diagonals are interpreted as m -digit binary numbers $w_{i1} \dots w_{im}$, $i = 0, 1, \dots, M$. Then the indices $1, \dots, M$ are associated with the diagonals as follows

$$\begin{aligned} 0 &\rightarrow 00 \dots 0, & 1 &\rightarrow 0 \dots 01, & 2 &\rightarrow 0 \dots 010, & 3 &\rightarrow 0 \dots 011, \dots, \\ M-2 &\rightarrow 11 \dots 10, & M &\rightarrow 11 \dots 11, \end{aligned} \quad (6)$$

where the number of models is $M + 1 = 2^m$. Standard theory of LS estimation with linear restrictions^{1,3} yields the restricted LS estimators

$$\hat{\beta}_i = \hat{\beta}_0 - \mathbf{QW}_i \hat{\theta} \quad (7)$$

for β under the models $\mathcal{M}_i, 0 \leq i \leq M$.

3. Model selection and averaging

The aim of model selection (MS) is to choose a model $\mathcal{M}_i, 0 \leq i \leq M$, from the set of candidate models $\mathcal{M}_0, \dots, \mathcal{M}_M$. Given an MS procedure S , the associated post MS estimator may be represented as

$$\hat{\beta}_S = \sum_{i=0}^M \mathbb{I}(S = \mathcal{M}_i) \hat{\beta}_i, \quad (8)$$

where $\hat{\beta}_i$ denotes the LS estimator of β under \mathcal{M}_i and $\mathbb{I}(\cdot)$ is the indicator function with the value 1 for the selected model and 0 for all other models. Akaike's information criterion AIC⁴ and Bayesian information criterion BIC,⁵ as well as the minimum description length (MDL) principle,⁶⁻⁸ for example, are well known MS criteria. However, traditionally by far the most common selection approach in practice is to carry out a sequence of tests in order to identify the nonzero regression coefficients and select the corresponding regressors. Forward selection, backward elimination, and stepwise regression are the best known examples of these techniques.⁹ It is not unusual that β is estimated from the selected model and the properties of the estimator are reported as if estimation had not been preceded by model selection. Deleting variables from a model increases bias and decreases variance. To minimize the mean square error (*MSE*) of estimation, a balance must be attained between the bias due to omitted variables and the variance due to parameter estimation.

Model averaging (MA) offers a more general way of weighting models than just by means of indicator functions like in model selection (8). Let $\lambda = (\lambda_0, \lambda_1, \dots, \lambda_M)'$ be a vector of nonnegative weights which sum to one

and thus $\boldsymbol{\lambda}$ lies on the \mathbb{R}^{M+1} unit simplex

$$\Delta^{M+1} = \{\boldsymbol{\lambda} \in [0, 1]^{M+1} : \sum_{i=0}^M \lambda_i = 1\}. \quad (9)$$

Then a model averaging LS estimator for $\boldsymbol{\beta}$ takes the form

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &= \sum_{i=0}^M \lambda_i \hat{\boldsymbol{\beta}}_i = \sum_{i=0}^M \lambda_i (\hat{\boldsymbol{\beta}}_0 - \mathbf{Q}\mathbf{W}_i \hat{\boldsymbol{\theta}}) \\ &= \hat{\boldsymbol{\beta}}_0 - \mathbf{Q}\mathbf{W} \hat{\boldsymbol{\theta}}, \end{aligned} \quad (10)$$

where $\mathbf{W} = \sum_{i=0}^M \lambda_i \mathbf{W}_i$. Hansen¹⁰ shows that a LS model averaging estimator like (10) can achieve lower *MSE* than any individual estimator (7). Magnus et al.¹¹ introduced the LS model averaging estimator (10) and called it weighted-average LS (WALS).

Magnus et al.¹¹ assume that the weights $\lambda_0, \lambda_1, \dots, \lambda_M$ in (9) are random and they depend on least squares residuals $\mathbf{M}\mathbf{y}$, i.e.

$$\lambda_i = \lambda_i(\mathbf{M}\mathbf{y}), \quad i = 0, 1, \dots, M. \quad (11)$$

Note especially that $\hat{\boldsymbol{\theta}}$ is a function of $\mathbf{M}\mathbf{y}$. Similarly in model selection (8), $I(S = \mathcal{M}_i) = \lambda_i(\mathbf{M}\mathbf{y}) = 1$ for exactly one $i \in \{1, \dots, M\}$. Thus model selection is a special case of model averaging. Note that the selection matrices \mathbf{W}_i , $0 \leq i \leq M$ are nonrandom $m \times m$ diagonal matrices whereas \mathbf{W} is a random $m \times m$ diagonal matrix with diagonal elements

$$\mathbf{w} = (w_1, \dots, w_m)', \quad 0 \leq w_i \leq 1, \quad i = 1, \dots, m. \quad (12)$$

For example, when $m = 3$, we have $M+1 = 2^3$ models to compare. If we use the indexing given in (6), the diagonal elements of the selection matrices $\mathbf{W}_i, i = 0, 1, \dots, 7$ are

$$\begin{array}{cccc} 0 : 000 & 1 : 001 & 2 : 010 & 3 : 011 \\ 4 : 100 & 5 : 101 & 6 : 110 & 7 : 111 \end{array}$$

and hence the diagonal entries of \mathbf{W}

$$w_1 = \lambda_4 + \lambda_5 + \lambda_6 + \lambda_7, \quad w_2 = \lambda_2 + \lambda_3 + \lambda_6 + \lambda_7, \quad w_3 = \lambda_1 + \lambda_3 + \lambda_5 + \lambda_7.$$

are random variables such that $0 \leq w_i \leq 1$, $i = 1, 2, 3$.

The equivalence theorem of Danilov and Magnus¹² provides a useful representation for the expectation, variance and *MSE* of the WALS estimator $\tilde{\boldsymbol{\beta}}$ given in (10). The theorem was proved under the assumptions

that the disturbances $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. $N(0, \sigma^2)$ and the weight vector λ satisfies the regularity conditions (9) and (11). By the theorem

$$E(\tilde{\beta}) = \beta - \mathbf{Q} E(\mathbf{W}\hat{\theta} - \theta), \quad \text{Var}(\tilde{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{Q}[\text{Var}(\mathbf{W}\hat{\theta})]\mathbf{Q}'$$

and hence

$$MSE(\tilde{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{Q}[MSE(\mathbf{W}\hat{\theta})]\mathbf{Q}'.$$

The major ingredient of the proof is that the estimator $\hat{\theta}$ in (6) and $\hat{\beta}_0$ are uncorrelated and under the normality assumption they are independent. Now the relatively simple estimator $\mathbf{W}\hat{\theta}$ of θ characterizes the important features of the more complicated WALS estimator $\tilde{\beta}$ of β .

There is a growing literature on MA, see Hoeting et al.¹³ for a review of Bayesian methods, and Claeskens and Hjort¹⁴ on frequentist methods. Hansen¹⁰ and Hansen and Racine,¹⁵ for example, have developed methods to estimate the model weights in view of reducing estimation variance while controlling omitted variables bias. In practice the number of weights to be estimated can be huge, and therefore the set of candidate models is usually restricted to a small fraction of all possible models. However, the effect of this "preselection" is usually ignored.

We assume the approach proposed by Magnus et al.¹¹ where instead of every single weight λ_i we estimate the diagonal elements (12) of \mathbf{W} . Then the core of the WALS estimator (10) will be to find a good shrinkage estimator

$$\tilde{\theta} = \mathbf{W}\hat{\theta}, \quad 0 \leq |\tilde{\theta}_i| \leq |\hat{\theta}_i|, \quad i = 1, \dots, m, \quad (13)$$

of θ . Magnus et al.¹¹ assumed that each diagonal element $w_j = w_j(\hat{\theta}_j)$ depends only on $\hat{\theta}_j$, the j th element of $\hat{\theta}$, $1 \leq j \leq m$. Since $\hat{\theta}_1, \dots, \hat{\theta}_m$ are independent under the normality assumption, also w_1, \dots, w_m are independent. Assuming that σ^2 is known, we have to find the best estimator of θ_j when $\hat{\theta}_j \sim N(\theta_j, \sigma^2)$, $1 \leq j \leq m$. Thus we have m independent estimation problems. The case of unknown σ^2 will be discussed later. If the number of auxiliary regressors is large, say $m = 50$, then computing time of WALS is only of order 50. If estimation of every single weight λ_i , $0 \leq i \leq M$ is required, the computing time will be of order 2^{50} . Thus the proposed WALS technique is computationally superior to techniques that require the estimation of every single weight.

4. Shrinkage with penalized LS

4.1. Shrinkage estimation

The essence of WALs estimation is the shrinkage estimator (13) of θ presented in (10), where $\hat{\theta}$ is the LS estimator of θ and \mathbf{W} is a random $m \times m$ diagonal matrix with diagonal elements w_i , $0 \leq w_i \leq 1$, $i = 1, \dots, m$ (see (12)). Thus w_i 's shrink the LS estimates $\hat{\theta}_i$ towards zero, and consequently $0 \leq |\tilde{\theta}_i| \leq |\hat{\theta}_i|$, $i = 1, \dots, m$. Further, we assume that the shrinkage functions are even: $w_i(-\hat{\theta}_i) = w_i(\hat{\theta}_i)$, $i = 1, \dots, m$. Thus the functions $\tilde{\theta}_i$ are odd: $\tilde{\theta}_i(-\hat{\theta}_i) = -\tilde{\theta}_i(\hat{\theta}_i)$. Magnus et al.¹¹ and Einmahl et al.²⁸ adopted a Bayesian view on estimation deciding on to advocate the Laplace and Subbotin estimators which are of shrinkage type. The Laplace and Subbotin estimators are defined in Subsection 4.4.

The proposed estimators (13) are computationally superior to estimators that require estimation of every single weight in (9), since in estimation of $\tilde{\theta}$ in (13) we have only m independent estimation problems $\tilde{\theta}_i = w_i \hat{\theta}_i$. We are now ready to define an important class of shrinkage estimators for θ . In the sequel \mathcal{S} denotes this class and we call the estimators in \mathcal{S} simply shrinkage estimators.

Definition 4.1. A real valued estimator δ of θ defined on \mathbb{R} is a shrinkage estimator if the following four conditions hold:

- (a) $0 \leq \delta(\hat{\theta}) \leq \hat{\theta}$ for $\hat{\theta} \geq 0$,
- (b) $\delta(-\hat{\theta}) = -\delta(\hat{\theta})$,
- (c) $\delta(\hat{\theta})/\hat{\theta}$ is nondecreasing on $[0, \infty)$ and
- (d) $\delta(\hat{\theta})$ is continuous,

where $\hat{\theta}$ is the LS estimator of θ .

In addition to shrinkage property (a) and antisymmetry (b), the definition puts two further requirements for shrinkage estimators. Consider now the condition (c). Denote $w(\hat{\theta}) = \delta(\hat{\theta})/\hat{\theta}$ for $\hat{\theta} > 0$ and think $\delta(\hat{\theta})$ as a weighted average of $\hat{\theta}$ and 0: $\delta(\hat{\theta}) = w(\hat{\theta})\hat{\theta} + (1 - w(\hat{\theta}))0$. The larger is $|\hat{\theta}|$, the better $\hat{\theta}$ is an estimator of θ . Hence, when $\hat{\theta}$ increases we wish to put more weight on $\hat{\theta}$ than on 0, i.e., we wish to make $w(\hat{\theta})$ larger. Thus we see that the condition (c) makes sense. Condition (d) is a minimal smoothness condition which guarantees certain stability of estimation in the sense that small changes of data cannot create excessive variation of estimates.

4.2. Penalized LS estimation

Fitting the orthogonalized model (2) can be considered as a two-step least squares procedure.¹ The first step is to calculate $\hat{\beta}_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and replace \mathbf{y} by $\mathbf{y} - \mathbf{X}\hat{\beta}_0 = \mathbf{M}\mathbf{y}$, where \mathbf{M} is defined in (3). Then denote $\mathbf{z} = \mathbf{U}'\mathbf{y}$, and note that from the definition of \mathbf{U} in (4) follows the equality $\mathbf{U}'\mathbf{M} = \mathbf{U}'$. Then the model \mathcal{M}_M in (5) takes the form

$$\mathbf{z} = \boldsymbol{\theta} + \mathbf{U}'\boldsymbol{\varepsilon}, \quad \mathbf{U}'\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\mathbf{I}_m). \quad (14)$$

The second step is to estimate $\boldsymbol{\theta}$ from the model (14).

In estimation of $\boldsymbol{\theta}$ we will use the penalized LS technique. If the penalty function satisfies proper regularity conditions, then the penalized LS yields a solution which is a shrinkage estimator of $\boldsymbol{\theta}$. In this approach we choose a suitable penalty function in order to get a shrinkage estimator with good risk properties. The related Bayesian technique is to impose certain restrictions on the prior density, see e.g. Einmahl et al.²⁸ So, we are able to characterize a variety of interesting estimators from which many have already shown their potential in applications. This technique is also computationally efficient.

The penalized least squares estimate (PenLS) of $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)'$ is the minimizer of

$$\frac{1}{2} \sum_{i=1}^m (z_i - \theta_i)^2 + \sum_{i=1}^m p_\lambda(|\theta_i|), \quad (15)$$

where $\lambda > 0$. It is assumed that the penalty function $p_\lambda(\cdot)$ is

- (i) nonnegative,
 - (ii) nondecreasing and
 - (iii) differentiable on $[0, \infty)$.
- (16)

Minimization of (15) is equivalent to minimization componentwise. Thus we may simply minimize

$$l(\theta) = \frac{1}{2}(z - \theta)^2 + p_\lambda(|\theta|) \quad (17)$$

with respect to θ .

Example 4.1. There are close connections between the PenLS and variable selection or the PenLS and ridge regression, for example. Taking the L_2 penalty $p_\lambda(|\theta|) = \frac{\lambda}{2}|\theta|^2$ yields the ridge estimator

$$\check{\theta}_R = \frac{1}{1 + \rho} z,$$

where $\rho > 0$ depends on λ . The hard thresholding penalty function

$$p_\lambda(|\theta|) = \lambda^2 - \frac{1}{2}(|\theta| - \lambda)^2 \mathbb{I}(|\theta| < \lambda)$$

yields the hard thresholding rule

$$\check{\theta}_H = z \{ \mathbb{I}(|z| > \lambda) \}, \quad (18)$$

where $\mathbb{I}(\cdot)$ is the indicator function. Then the minimizer of the expression (15) is $z_j \{ \mathbb{I}(|\theta_j| > \lambda) \}$, $j = 1, \dots, m$, and it coincides with the best subset selection for orthonormal designs. In statistics (see e.g. Morris et al.¹⁶) and in econometrics (see, e.g. Judge *et al.*¹⁷), the hard thresholding rule is traditionally called the pretest estimator.

The following theorem gives sufficient conditions for the PenLS estimate $\check{\theta}$ of θ to be a shrinkage estimator. Further, the theorem provides the lower bound of the mean squared error

$$MSE(\theta, \check{\theta}) = E[\check{\theta}(z) - \theta]^2 = \text{Var}[\check{\theta}(z)] + \text{Bias}(\theta, \check{\theta}),$$

where $\text{Bias}(\theta, \check{\theta}) = \{E[\check{\theta}(z)] - \theta\}^2$. This lower bound is called the *efficiency bound*.

Theorem 4.1. *We assume that the penalty function $p_\lambda(\cdot)$ satisfies the assumptions (16). We make two assertions.*

(i) *If the three conditions hold*

- (1) *the function $-\theta - p'_\lambda(\theta)$ is strictly unimodal on $[0, \infty)$,*
- (2) *$p'_\lambda(\cdot)$ is continuous and nonincreasing on $[0, \infty)$, and*
- (3) *$\min_\theta \{|\theta| + p'_\lambda(|\theta|)\} = p'_\lambda(0)$,*

then the PenLS estimate $\check{\theta}$ of θ belongs to the shrinkage family \mathcal{S} .

(ii) *If the conditions of the assertion (i) hold and z follows the normal distribution $N(0, \sigma^2)$, where σ^2 is known, the efficiency bound of $\check{\theta}$ is*

$$\inf_{\check{\theta} \in \mathcal{S}} MSE(\theta, \check{\theta}) = \frac{\theta^2}{1 + \theta^2}.$$

Proof. (i) The derivative $l'(\theta)$ of the function $l(\theta)$ to be minimized in (17) is

$$l'(\theta) = \text{sgn}(\theta) \{ |\theta| + p'_\lambda(|\theta|) \} - z.$$

If the three conditions in (i) hold, then by Theorem 1 in Antoniadis and Fan¹⁹ the solution to the minimization problem (17) exists, is unique and takes the form

$$\check{\theta}(z) = \begin{cases} 0, & \text{if } |z| \leq p_0, \\ z - \text{sgn}(z) p'_\lambda(|z|), & \text{if } |z| > p_0, \end{cases} \quad (19)$$

where $p_0 = \min_{\theta \geq 0} \{\theta + p'_\lambda(\theta)\}$. Clearly the solution (19) is antisymmetric, i.e. $\check{\theta}(-z) = -\check{\theta}(z)$. Since $p'_\lambda(z) \geq 0$ for $z \geq 0$, $\check{\theta}(z)$ satisfies the shrinkage property (a) of definition 4.1: $0 \leq \check{\theta}(z) \leq z$ for $z \geq 0$.

If $\min_{\theta} \{|\theta| + p'_\lambda(|\theta|)\} = p'_\lambda(0)$, then $p_0 = p'_\lambda(0)$ and the PenLS estimator (19) is continuous. Furthermore, since $p'_\lambda(\cdot)$ is nonincreasing on $[0, \infty)$, it follows that $\check{\theta}(z)/z$ defined by (19) is nondecreasing on $[0, \infty)$. Hence the estimator (19) fulfils the condition (c) in Definition 4.1. Thus we have proved that the PenLS estimator (19) belongs to the shrinkage class \mathcal{S} .

(ii) By the assertion (i) the PenLS estimator $\check{\theta}(z)$ belongs to shrinkage family \mathcal{S} , and consequently $\check{\theta}(z)$ satisfies the regularity conditions R1 in Magnus.²⁰

- (a) $0 \leq \check{\theta}(z)/z \leq 1$ for all z ,
- (b) $\check{\theta}(-z)/(-z) = \check{\theta}(z)/z$ for all z ,
- (c) $\check{\theta}(z)/z$ is nondecreasing on $[0, \infty)$ and
- (d) $\check{\theta}(z)/z$ is continuous.

Hence by Theorem A7 in Magnus²⁰ the efficiency bound for the shrinkage estimators \mathcal{S} is

$$\inf_{\check{\theta} \in \mathcal{S}} MSE(\theta, \delta) = \frac{\theta^2}{1 + \theta^2}.$$

This concludes the proof of the theorem. \square

Note that the pretest estimator $\check{\theta}_H$ given in (18) is not continuous, and hence it does not belong to the class of shrinkage estimators \mathcal{S} . Magnus²¹ demonstrates a number of undesirable properties of the pretest estimator. It is inadmissible and there is a range of values for which the MSE of $\check{\theta}_H$ is greater than the MSE of both the least squares estimator $\hat{\theta}(z) = z$ and the null estimator $\hat{\theta}(z) \equiv 0$. The traditional pretest at the usual 5% level of significance results in an estimator that is close to having worst possible performance with respect to the MSE criterion in the neighborhood of the value $|\theta/\sigma| = 1$ which was shown to be of crucial importance.

Example 4.2. The L_q penalty $p_\lambda(|\theta|) = \lambda |\theta|^q$, $q \geq 0$ results in a bridge regression.²² The derivative $p'_\lambda(\cdot)$ of the L_q penalty is nonincreasing on

$[0, \infty)$ only when $q \leq 1$ and the solution is continuous only when $q \geq 1$. Therefore, only L_1 penalty in this family yields a shrinkage estimator. This estimator is a soft thresholding rule, proposed by Donoho and Johnstone,²³

$$\tilde{\theta}_S = \text{sgn}(z)(|z| - \lambda)_+, \quad (20)$$

where z_+ is shorthand for $\max\{z, 0\}$. LASSO²⁴ is the PenLS estimate with the L_1 penalty in the general least squares and likelihood settings.

Since we have the efficiency bound of the PenLS estimators (19), the *regret* of $\tilde{\theta}(z)$ can be defined as

$$r(\theta, \tilde{\theta}) = \text{MSE}(\theta, \tilde{\theta}) - \frac{\theta^2}{1 + \theta^2}.$$

We wish to find an estimator with the desirable property that its risk is uniformly close to the infeasible efficiency bound. In search of such an estimator we may adopt the minimax regret criterion where we minimize the maximum regret instead of the maximum risk. An estimator $\tilde{\theta}^*$ is *minimax regret* if

$$\sup_{\theta} r(\theta, \tilde{\theta}^*) = \inf_{\tilde{\theta} \in \mathcal{S}} \sup_{\theta} r(\theta, \tilde{\theta}).$$

In theoretical considerations σ^2 is assumed to be known, and hence we can always consider the variable z/σ . Then expectation E is simply taken with respect to the $N(\theta, 1)$ distribution, and comparison of estimators risk performance is done under this assumption. In practical applications we replace the unknown σ^2 with s^2 , the estimate in the unrestricted model. Danilov¹² demonstrated that effects of estimating σ^2 are small in case of Laplace estimator. We expect the approximation to be accurate for other shrinkage estimators too, although more work is needed to clarify this issue.

4.3. Good PenLS shrinkage estimators

In this subsection we consider properties of three well known PenLS estimators which are shrinkage estimators. The performance of two of them is also displayed in simulation experiments. Bruce and Gao²⁵ compared hard and soft thresholding rules and showed that hard thresholding tends to have bigger variance whereas soft thresholding tends to have bigger bias. To remedy the drawbacks of hard and soft thresholding, Fan and Li¹⁸ suggested using continuous differentiable penalty function defined by

$$p'_\lambda(|\theta|) = \lambda \{\mathbb{I}(|\theta| \leq \lambda) + \frac{(a\lambda - |\theta|)_+}{(a-1)\lambda} \mathbb{I}(|\theta| > \lambda)\} \quad (21)$$

for some $a > 2$ and $\theta > 0$. The penalty (21) is called *smoothly clipped absolute deviation* (SCAD) penalty. Note that if the penalty function in (15) is constant, i.e. $p'(|\theta|) = 0$, then the rule in (19) takes the form $\hat{\theta}(z) \equiv z$ which is unbiased. Since the SCAD penalty $p'_\lambda(\theta) = 0$ for $\theta > a\lambda$, the resulting solution (Fan and Li¹⁸)

$$\check{\theta}_{scad}(z) = \begin{cases} \text{sgn}(z)(|z| - \lambda)_+, & \text{if } |z| \leq 2\lambda, \\ \frac{(a-1)z - \text{sgn}(z)a\lambda}{(a-2)}, & \text{if } 2\lambda < |z| \leq a\lambda, \\ z, & \text{if } |z| > a\lambda \end{cases} \quad (22)$$

tends to be unbiased for large values of z . This estimator (22) can be viewed as a combination of soft thresholding for "small" $|z|$ and hard thresholding for "large" $|z|$, with a piecewise linear interpolation inbetween.

The SCAD estimator is closely related to the firm thresholding rule of Bruce and Gao:²⁵

$$\check{\theta}_F(z) = \begin{cases} 0, & \text{if } |z| \leq \lambda_1, \\ \text{sgn}(z) \frac{\lambda_2(|z| - \lambda_1)}{\lambda_2 - \lambda_1}, & \text{if } \lambda_1 < |z| \leq \lambda_2, \\ z, & \text{if } |z| > \lambda_2, \end{cases} \quad (23)$$

where $0 < \lambda_1 < \lambda_2$. This rule was also suggested to ameliorate the drawbacks of hard and soft thresholding. For soft thresholding $p'(|\theta|) = \lambda$ for all θ , and $\hat{\theta}_S$ is biased also for large values of $|z|$. Bruce and Gao²⁵ showed that $MSE(\theta, \hat{\theta}_S) \rightarrow 1 + \lambda^2$ as $\theta \rightarrow \infty$ whereas $MSE(\theta, \hat{\theta}_F) \rightarrow 1$ as $\theta \rightarrow \infty$ (Bruce and Gao²⁵) when $\lambda_2 < \infty$.

Breiman²⁶ applied the non-negative garrote rule

$$\check{\theta}_G(z) = \begin{cases} 0, & \text{if } |z| \leq \lambda, \\ z - \lambda^2/z, & \text{if } |z| > \lambda \end{cases} \quad (24)$$

to subset selection in regression to overcome the drawbacks of stepwise variable selection rule and ridge regression. The MSE for the estimator $\hat{\theta}_G$ is comparable to that for the firm thresholding rule.^{25,27} It is straightforward to show that the soft thresholding (20), SCAD (22), firm thresholding (23) and non-negative garrote (24) estimators belong to the shrinkage class \mathcal{S} (Definition 4.1). The usual LS estimator $\hat{\theta}(z) \equiv z$ is a good candidate for large z , and hence we wish that for large z an estimator $\check{\theta}(z)$ is close to z in the sense that $z - \check{\theta}(z)$ converges to zero. It can be readily seen that the estimators $\check{\theta}_{scad}$, $\check{\theta}_F$ and $\check{\theta}_G$ have this property, i.e. $z - \check{\theta}(z) \rightarrow 0$ as $z \rightarrow \infty$ when $\check{\theta}(z)$ is any of the foregoing three estimators. For the soft thresholding rule $z - \check{\theta}_S(z)$ converges to a positive constant, but not to zero.

4.4. The Laplace and Subbotin estimators

Magnus²⁰ addressed the question of finding an estimator of θ which is admissible, has bounded risk, has good risk performance around $\theta = 1$, and is optimal or near optimal in terms of minimax regret when $z \sim N(\theta, 1)$. The Laplace estimator

$$\hat{\theta}_L(z) = z - h(y)c$$

proved to be such an estimator, when $c = \log 2$ and $h(\cdot)$ is a given antisymmetric monotonically increasing function on $(-\infty, \infty)$ with $h(0) = 0$ and $h(\infty) = 1$. The Laplace estimator is the mean of the posterior distribution of $\theta|z$ when a Laplace prior for θ with $\text{median}(\theta) = 0$ and $\text{median}(\theta^2) = 1$ is assumed. In search of prior which appropriately reflects the notion of ignorance, Einmahl et al.²⁸ arrived at the Subbotin prior that belongs to the class of reflected gamma densities. In practical applications they recommended the Subbotin prior

$$\pi(\theta) = \frac{c^2}{4} e^{-c|\theta|^{1/2}}$$

with $c = 1.6783$ which should stay close to the Laplace prior. Magnus et al.¹¹ and Einmahl et al.²⁸ also showed that the computational burden of the Laplace and Subbotin estimators is light when applied in the context of weighted average least squares (WALS). In our simulation experiments we compare the performance of these two Bayesian estimators, the Laplace and Subbotin, with the performance of the penalized LS estimators.

4.5. Implementation using penalized LS

We now recap the main steps of the penalized LS estimation of the parameters β and γ in the model (1). To orthogonalize the model (1) fix a matrix \mathbf{C} such that $\mathbf{C}'\mathbf{Z}'\mathbf{M}\mathbf{Z}\mathbf{C} = \mathbf{I}_m$. We can use the spectral decomposition $\mathbf{Z}'\mathbf{M}\mathbf{Z} = \mathbf{P}\mathbf{\Phi}\mathbf{P}'$ of $\mathbf{Z}'\mathbf{M}\mathbf{Z}$ to have $\mathbf{C} = \mathbf{P}\mathbf{\Phi}^{-1/2}$, where $\mathbf{\Phi} = \text{diag}(\phi_1, \dots, \phi_m)$ is the diagonal matrix of the eigenvalues of $\mathbf{Z}'\mathbf{M}\mathbf{Z}$ and the columns of \mathbf{P} is an orthonormal set of eigenvectors corresponding to these eigenvalues.

- (1) Compute $\mathbf{y} - \mathbf{X}\hat{\beta}_0 = \mathbf{M}\mathbf{y}$ and $\hat{\boldsymbol{\theta}} = \mathbf{C}'\mathbf{Z}'\mathbf{M}\mathbf{y}$, where $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.
- (2) Compute $\hat{\boldsymbol{\theta}} := \hat{\boldsymbol{\theta}}/\sigma$, where $\boldsymbol{\theta}$ denotes $\boldsymbol{\theta}/\sigma$ and σ^2 is assumed to be known.
- (3) For $j = 1, \dots, m$ compute the PenLS estimate $\check{\theta}_j$ and its variance $\check{\omega}_j^2$. Denote $\check{\boldsymbol{\theta}} = (\check{\theta}_1, \dots, \check{\theta}_m)'$ and $\check{\boldsymbol{\Omega}} = \text{diag}(\check{\omega}_1^2, \dots, \check{\omega}_m^2)$.

(4) The PenLS estimates for γ and β are

$$\tilde{\gamma} = \sigma \mathbf{C} \check{\underline{\theta}} \quad \text{and} \quad \check{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'(\mathbf{y} - \mathbf{Z}\tilde{\gamma}),$$

since $\gamma = \mathbf{C}\theta$ and $\theta = \sigma\underline{\theta}$.

(5) The variance for $\tilde{\gamma}$ and $\check{\beta}$ are

$$\begin{aligned} \text{Var}(\tilde{\gamma}) &= \sigma^2 \mathbf{C} \check{\underline{\Omega}} \mathbf{C}' \\ \text{Var}(\check{\beta}) &= \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} + \underline{\mathbf{Q}} \text{var}(\tilde{\gamma}) \underline{\mathbf{Q}}', \end{aligned}$$

where $\underline{\mathbf{Q}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Z}$. Finally we have $\text{Cov}(\check{\beta}, \tilde{\gamma}) = -\underline{\mathbf{Q}} \text{var}(\tilde{\gamma})$.

In practice σ^2 is unknown and it is replaced with s^2 , the sample variance estimated in the unrestricted model.

5. The costs of initial hospitalization for a first hip fracture

We compare the estimation techniques presented in this paper on hip fracture data. The original purpose of our dataset is to compare treatment costs of hip fracture patients between hospital districts in Finland. In this paper we use it to demonstrate the performance of various penalized least squares estimators.

The dataset was obtained by combining data from several national registries.²⁹ The costs of the first institutionalization period of first time hip fracture patients in Finland were calculated in the time period of 1999 – 2005. There are a total of 21 hospital districts in Finland, but in the estimations in this paper we are only using the seven largest districts. The dataset was made more homogenous by keeping such patients in the data who had not been institutionalized before the fracture and who were not institutionalized after the fracture either. Patients who died within a year after the fracture were removed. The final dataset used in this paper contained 11961 patients of age 50 or older.

As the dependent variable in our model we are using the cost of the first continuous institutionalization period. In our model we have 7 focus regressors, which are dummy variables for the six largest hospital districts and 31 auxiliary regressors. The largest hospital district was taken as the baseline. The set of auxiliary regressors contains information on the patients such as gender, age and time between fracture and operation and a number of important comorbidities like congestive heart failure, diabetes and cancer. The auxiliary regressors are intended to reflect the mix of patients treated in a hospital district.

6. Simulation experiments

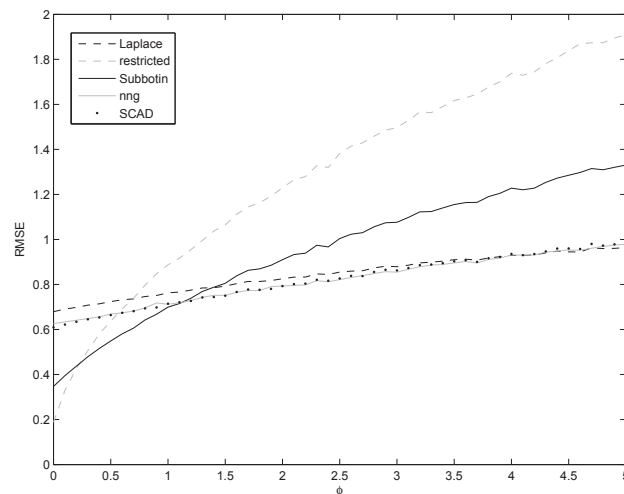


Fig. 1. The RMSE values of the Laplace, non-negative garrote (nng), restricted LS, SCAD and Subbotin estimators are compared to the unrestricted model. $RMSE = 1$ is the $RMSE$ -value of the unrestricted model.

The purpose of the simulations is to compare the performance of various PenLS estimators, including the restricted LS estimator, with the performance of the Laplace estimator within a realistic set-up. The Laplace estimator has been shown to be theoretically and practically superior to many existing MA methods.^{11,20} Recently Einmahl et al.²⁸ proposed a competitor for it, the Subbotin estimator. Therefore also the Subbotin estimator is included in our simulation study.

We use the LS estimator in the unrestricted model as our benchmark. We take the estimates from the unrestricted model as the 'true' parameter values. We do not generate the disturbances from a theoretical distribution, but the disturbances are obtained by resampling the LS residuals of the estimated unrestricted model. Thus the simulations are based on real data, not on generated data. The disturbances in each round of the simulation experiment are obtained by randomly selecting 2000 numbers with replacement from the LS residuals. In order to gain broader perception of

the estimators performance we use different values of γ by scaling it. This is carried out so that we replace γ by $\tau\gamma$ where the scale factor τ is obtained from the equality

$$\phi = \tau^2 \gamma' \mathbf{Z}' \mathbf{M} \mathbf{Z} \gamma,$$

when we let ϕ vary between 0 and 5. Here ϕ can be considered the approximation of the theoretical F-ratio $\gamma' \mathbf{Z}' \mathbf{M} \mathbf{Z} \gamma / (m\sigma^2)$.

We concentrate on the performance of our focus parameters, the β -parameters. Based on 10000 replications we approximate the distribution of $\hat{\beta}$. The estimators are evaluated in terms of the root mean squared error (*RMSE*). Let $\hat{\beta}^{(i)}$ denote the estimate of β in the i -th replication, and we compute

$$RMSE(\hat{\beta}) = \sqrt{\sum_{i=1}^N \|\hat{\beta}^{(i)} - \beta\|^2 / N},$$

where $\|\cdot\|$ denotes the Euclidean norm, $N = 10000$ is the number of replicates and β is the estimate from the unrestricted model. The *RMSE* of each estimator is computed. Since the LS estimator in the unrestricted model is used as the benchmark, the *RMSE* of an estimator is divided by the *RMSE* computed from the unrestricted model. So, *RMSE* = 1 in In Figure 1 means that the *RMSE* of an estimator is equal to that of the unrestricted model.

The parameter values of the SCAD and the non-negative garrote were chosen so that the theoretical risk (*MSE*) of the estimators are uniformly close to the efficiency bound of the shrinkage estimators. For SCAD we used parameter values $a = 5$ and $\lambda = 0.5$ and for the non-negative garrote we take $\lambda = 0.01$. For these parameter values the *MSE* of the SCAD and the non-negative garrote were also close to the *MSE* of the Laplace estimator.

In Figure 1 we have compared the *RMSE*'s of the competing estimators as ϕ increases. We observe that the Laplace estimator and SCAD perform better than the unrestricted model with all ϕ values. The SCAD estimator does a little better than Laplace with small and intermediate ϕ values. The non-negative garrote estimator performs equally well with SCAD. Subbotin performs very well with $\phi < 1$, but with larger ϕ values loses to SCAD, Laplace, non-negative garrote and the unrestricted model.

7. Concluding remarks

In model selection one attempts to use the data to find a single "winning" model, according to a given criterion, whereas with model averaging (MA)

one seeks a smooth compromise across a set of competing models. Most existing MA methods are based on estimation of all model weights using exponential Akaike information criterion (AIC) or Bayesian information criterion (BIC) weights, for example. A common challenge for a regression analyst is the selection of the best subset from a set of m predictor variables in terms of some specified criterion. Then the number of competing models is 2^m , and consequently the computational burden to estimate all the model weights becomes soon too heavy when m is large.

It turns out, that the quality of the WALS (10) estimator depends on the shrinkage estimator of the auxiliary parameter γ where each shrinkage factor is a sum of model weights. So, estimation of 2^m model weights is converted into estimation of m shrinkage factors with trivial computational burden. We define the class of shrinkage estimators in view of MA and show that these shrinkage estimators can be constructed by putting appropriate restrictions on the penalty function. Utilizing the relationship between shrinkage and parameter penalization, we are able to build up computationally efficient MA estimators which are easy to implement into practice. These estimators include some known recent contributions, like the non-negative garrote of Breiman,²⁶ the lasso-type estimator of Tibshirani²⁴ and the SCAD estimator of Fan and Li.¹⁸ In the simulation experiments we assess the quality of an estimator in terms of its *RMSE*. In this competition the winners were the SCAD and non-negative garrote but the Laplace estimator did almost as well.

References

1. G. A. F. Seber, *Linear Regression Analysis* (Wiley, New York, 1977).
2. R. A. Horn and C. R. Johnson, *Matrix Analysis* (Cambridge University Press, Cambridge, 1985).
3. G. A. F. Seber and A. J. Lee, *Linear Regression Analysis*, 2nd edn. (Wiley, New York, 2003).
4. H. Akaike, Information theory as an extension of the maximum likelihood principle. Pages 267–281 in B. N. Petrov, and F. Csaki, eds. (Second International Symposium on Information Theory, Akademiai Kiado, Budapest, 1973).
5. G. Schwarz, Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464 (1978).
6. J. Rissanen, Modeling by Shortest Data Description. *Automatica*, 14, No. 1, 465–471 (1978).
7. J. Rissanen, *Information and Complexity in Statistical Modeling*. (Springer, New York, 2007).
8. J. Rissanen, *Optimal Parameter Estimation*. (Cambridge University Press, Cambridge, 2011).

9. A. Miller, *Subset Selection in Regression*. (Chapman & Hall, Boca Raton, 2002).
10. B. E. Hansen Least squares model averaging. *Econometrika*, 75, 1175–1189 (2007).
11. J. R. Magnus, O. Powell, and P. Prüfer, A comparison of two model averaging techniques with an application to growth empirics. *Journal of Econometrics*, 154, 139–153 (2010).
12. D. Danilov, and J. R. Magnus, On the harm that ignoring pretesting can cause. *Journal of Econometrics*, 122, 27–46 (2004).
13. J. A. Hoeting, D. Madigan, A. E. Raftery and C. T. Volinsky, Bayesian model averaging: A tutorial (with discussion). *Statistical Science*, 14, 382–417 (1999).
14. N. I. Hjort, and G. Claeskens, Frequentist model averaging estimators. *Journal of the American statistical Association*, 98, 879–899 (2003).
15. B. E. Hansen and J. Racine, Jackknife model averaging. *Journal of Econometrics*, forthcoming (2011).
16. C. Morris, R. Radhakrishnan and S. L. Sclove, Nonoptimality of preliminary test estimators for the mean of a multivariate normals distribution. *Annals of Mathematical Statistics*, 43, 1481–1490 (1972).
17. G. G. Judge, W. E. Griffiths, R. C. Hill, H. Lutkepohl and T. C. Lee, *The Theory and Practice of Econometrics*, (Wiley, New York, 1985).
18. J. Fan and R. Li, Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American statistical Association*, 96, 1348–1360 (2001).
19. A. Antoniadis and J. Fan, Regularization of Wavelets Approximations. *Journal of the American statistical Association*, 96, 939–967 (2001).
20. J. R. Magnus, Estimation of the mean of a univariate normal distribution with a known variance. *Econometrics Journal*, 5, 225–236 (2002).
21. J. R. Magnus, The traditional pretest estimator. *Theory of Probability and Its Applications*, 44, 293–308 (1999).
22. I. E. Frank and J. H. Friedman, A statistical view of some chemometrics regression tools. *Technometrics*, 35, 109–148 (1993).
23. D. L. Donoho and I. M. Johnstone, Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81, 425–456 (1994).
24. R. Tibshirani, Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society B*, 1, 267–288 (1996).
25. A. G. Bruce and H.-Y. Gao, Understanding WaveShrink: Variance and bias estimation. *Biometrika*, 83, 727–745 (1996).
26. L. Breiman, Better subset regression using nonnegative garrote. *Technometrics*, 37, 373–384 (1995).
27. H.-Y. Gao, Wavelet Shrinkage Denoising Using the Non-Negative Garrote. *Journal of Computational and Graphical Statistics*, 7, 469–488 (1998).
28. J. H. J. Einmahl, K. Kumar and J. R. Magnus Bayesian model averaging and the choice of prior. *CentER Discussion Paper*, No. 2011–003 (2011).
29. R. Sund, M. Juntunen, P. Lühje, T. Huusko, M. Mäkelä, M. Linna, A. Liski, U. Häkkinen, *PERFECT - Hip Fracture, Performance, Effectiveness and Cost of Hip Fracture Treatment Episodes* (In Finnish), National Research

[5] Liski, E. P. and Liski, A. (2008)

MDL model averaging for linear regression

In: Grünwald, P. et al. (eds.) *Festschrift in honor of Jorma Rissanen on the occasion of his 75th birthday*, Tampere, Finland TICSP series.

MDL Model Averaging for Linear Regression

ERKKI P. LISKI¹ and ANTTI LISKI²

University of Tampere¹ and Tampere University of Technology²
Finland

Abstract

Estimators formed after model selection really are like mixtures of many potential estimators. Sometimes it is advantageous to smooth estimators across several models, rather than rely only on the model that is suggested by a single selection criterion. The main theme of this paper is the problem of selecting the weights for averaging across estimates obtained from a set of models. Some existing model average (MA) methods are based on exponential AIC or BIC weights (e.g. Burnham and Anderson 2002). Bayesian model averaging is a related technique (see e.g. Hoeting et al. 1999). Recently Leung and Barron (2006) and Hansen (2007) have developed methods for combining estimators from various models. This paper considers selecting the model weights by using Rissanen's MDL criterion and compares the potential performance of alternative MA estimators in simulation experiments.

2000 *Mathematics Subject Classification.* 62B10, 62J05, 62F99.

Key words or phrases. Model selection, NML, AIC, BIC, Mallows' C_p .

1 Introduction

In statistical practice one typically has multiple plausible models available. Model selection is most often regarded as a way to select just the best model, and then inference is conditioned on that model. In regression a common practice is to decide which variables to include in the model, and to use these variables to fit the response. A large number of criteria has been developed over the past few decades to select the best model.

It is known that model selection procedures can be unstable, as a small perturbation in the data may lead to significant changes in model choice. If the inference done with an estimate on the chosen model does not take into account model uncertainty, it often means underreporting of variability. Model averaging (MA) is an alternative to model selection. There is a large Bayesian literature on MA, for literature reviews see e.g. Draper (1995) and Hoeting et al. (1999). Given a set of models, we may find several plausible models according to some model selection criterion. In this case, it has been suggested estimation strategies that utilize more than just a single model. This entails a weighted average estimator for many alternative models. Buckland et al. (1997) suggested exponential AIC and BIC weights (see also Burnham and Anderson 2002). Hjort and Claeskens (2003) developed a general large-sample likelihood apparatus for MA estimators.

The Minimum Description Length (MDL) principle provides a generic solution to the model selection problem. By viewing models as a means of providing statistical descriptions of observed data, the comparison between competing models is based on the stochastic complexity (SC) of each description. The Normalized Maximum Likelihood (NML) form of the SC (Rissanen 1996) contains a component that may be interpreted as the parametric complexity of the model class. Once the SC for the data, relative to a class of suggested models, is calculated, it serves as

a criterion for selecting the optimal model with the smallest SC. This is the MDL principle (Rissanen 1978, 1986, 1996, 2000, 2007) for model choice.

In this paper we consider the NML density as an implementation of the MDL principle for model selection in the linear regression context, where attention is restricted to Gaussian linear models. Then we propose a model average estimator with weights selected by the MDL criterion. It turns out that, under squared error loss, the resulting mixture estimator usually performs better than the corresponding selection based estimator.

2 The Model

We have n pairs of observations $(y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)$, where y_i is real valued and \mathbf{x}_i is a $k_M \times 1$ vector, $1 \leq i \leq n$. Assume that the data follow a classical nonparametric regression model

$$y_i = \mu(\mathbf{x}_i) + \sigma \varepsilon_i, \quad i = 1, \dots, n, \quad (1)$$

where $\varepsilon_1, \dots, \varepsilon_n$ are independent random variables such that for each $1 \leq i \leq n$

$$E(\varepsilon_i | \mathbf{x}_i) = 0 \quad \text{and} \quad E(\varepsilon_i^2 | \mathbf{x}_i) = 1$$

and the positive constant σ defines the scale of the additive error $\sigma \varepsilon_i$.

Model (1) is called a nonparametric regression model when μ belongs to some general (infinite dimensional) function class. Here we assume that μ is in the space of square integrable functions L_2 whose elements admit representations as infinite dimensional linear models for which

$$\mu(\mathbf{x}) = \sum_{j=1}^{\infty} \beta_j \varphi_j(\mathbf{x}) \quad (2)$$

for some set of known functions $\{\varphi_1, \varphi_2, \dots\}$ and real valued coefficients β_1, β_2, \dots . We assume that (2) converges in mean square, i.e.

$$E[\mu(\mathbf{x}) - \mu_m(\mathbf{x})]^2 \rightarrow 0 \quad \text{as} \quad m \rightarrow \infty,$$

where

$$\mu_m(\mathbf{x}) = \sum_{j=1}^m \beta_j \varphi_j(\mathbf{x}).$$

The practical significance of (2) is that any $\mu \in L_2$ may be well approximated by $\mu_m(\mathbf{x})$ with a finite number of m terms. In the sequel we denote generally $x_{ij} = \varphi_j(\mathbf{x}_i)$. Note that the above approach is a standard technique in nonparametric regression (see e.g. Efremovich 1999 and Eubank 1999). This approach is also similar to series estimators in econometrics (see e.g. Newey 1997).

Now the model (1) can be written as a linear model

$$y_i = \sum_{j \in \mathcal{M}_m} x_{ij} \beta_j + b_{im} + \sigma \varepsilon_i, \quad i = 1, 2, \dots, n, \quad (3)$$

where $\mathcal{M}_m = \{1, 2, \dots, k_m\}$ with $k_m \leq n$,

$$b_{im} = \sum_{j=k_m+1}^{\infty} \beta_j x_{ij}$$

is the approximation error and the random errors $\varepsilon_1, \dots, \varepsilon_n$ are like in (1). Here the quantity k_m plays the role of a smoothing parameter. Sometimes $\mu_m(\mathbf{x})$ is called a truncated series approximation of μ and k_m a truncation point.

To obtain an estimate of μ one may employ an approximating linear model by omitting b_i which effect is considered negligible. In matrix notation an approximating model \mathcal{M}_m takes the form

$$\mathbf{y} = \boldsymbol{\mu}_m + \sigma \boldsymbol{\varepsilon}, \quad (4)$$

where $\mathbf{y} = (y_1, \dots, y_n)'$, $\boldsymbol{\mu}_m = \mathbf{X}_m \boldsymbol{\beta}_m$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)'$ and $\boldsymbol{\beta}_m = (\beta_1, \dots, \beta_{k_m})'$ is the $k_m \times 1$ vector of unknown regression coefficients. Here \mathbf{X}_m is the $n \times k_m$ matrix with ij element x_{ij} . We shall consider a set of approximating models $\{\mathcal{M}_1, \dots, \mathcal{M}_M\}$ such that $\mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq \dots \subseteq \mathcal{M}_M \subseteq \{1, 2, \dots, n\}$, where \mathcal{M}_m refers to the model (4). We suppose that M is an integer for which the matrix \mathbf{X}_{k_M} is of full column rank. Thus $k_1 \leq k_2 \leq \dots \leq k_M$, and consequently all \mathbf{X}_m with $1 \leq m \leq M$ are of full column rank.

Then the least-squares estimate of $\boldsymbol{\beta}_m$ is

$$\hat{\boldsymbol{\beta}}_m(\mathbf{y}) = (\mathbf{X}_m' \mathbf{X}_m)^{-1} \mathbf{X}_m' \mathbf{y},$$

and the corresponding estimate of $\boldsymbol{\mu}_m$ is

$$\hat{\boldsymbol{\mu}}_m = \mathbf{H}_m \mathbf{y}, \quad (5)$$

where \mathbf{H}_m denotes the projection matrix $\mathbf{X}_m (\mathbf{X}_m' \mathbf{X}_m)^{-1} \mathbf{X}_m'$. Often the regression literature refers to the matrix \mathbf{H}_m as the hat matrix. Denote $\mathbf{b}_m = (b_{1m}, \dots, b_{nm})'$ and note that $\boldsymbol{\mu} = \boldsymbol{\mu}_m + \mathbf{b}_m$. Thus $(\mathbf{I} - \mathbf{H}_m) \boldsymbol{\mu} = (\mathbf{I} - \mathbf{H}_m) \mathbf{b}_m$, and consequently $\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_m = (\mathbf{I} - \mathbf{H}_m) \mathbf{b}_m - \sigma \mathbf{H}_m \boldsymbol{\varepsilon}$. Therefore the model error $r_m = \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_m\|^2$ is

$$r_m = \mathbf{b}_m' (\mathbf{I} - \mathbf{H}_m) \mathbf{b}_m + \sigma^2 \boldsymbol{\varepsilon}' \mathbf{H}_m \boldsymbol{\varepsilon} - 2 \mathbf{b}_m' (\mathbf{I} - \mathbf{H}_m) \mathbf{H}_m \boldsymbol{\varepsilon}. \quad (6)$$

Taking the conditional expectation of r_m we obtain

$$E(r_m | \mathbf{x}_1, \dots, \mathbf{x}_n) = \mathbf{b}_m' (\mathbf{I} - \mathbf{H}_m) \mathbf{b}_m + \sigma^2 k_m,$$

since by assumptions of the model (1) the conditional expectation of the last term in the expression (6) is zero and $E(\sigma^2 \boldsymbol{\varepsilon}' \mathbf{H}_m \boldsymbol{\varepsilon} | \mathbf{x}_1, \dots, \mathbf{x}_n) = \sigma^2 k_m$.

Example. Assume that (2) is an orthogonal series representation for $\mu(x)$, $x \in [0, 1]$, where

$$\beta_j = \int_0^1 \varphi_j(x) \mu(x) dx, \quad j = 1, 2, \dots$$

and $\{1, \varphi_2, \varphi_3, \dots\}$ is an orthonormal basis for $\mu \in L_2[0, 1]$. An example of such a basis is

$$\varphi_j(x) = \sqrt{2} \cos((j-1)\pi x), \quad j = 1, 2, \dots \quad (7)$$

Other popular examples of orthogonal basis functions are orthogonal polynomials and wavelets. In this example we also assume that the basis functions are orthonormal with respect to the uniform design $x_j = (j-1/2)/n$, $j = 1, \dots, n$ like the cosine basis (7):

$$\sum_{i=1}^n \varphi_j(x_i) \varphi_k(x_i) = \begin{cases} 0, & j \neq k \\ n, & j = k \end{cases} \quad (8)$$

for all $j, k \in \{1, 2, \dots\}$ and $\varphi_1 \equiv 1$.

Using the orthogonality properties (8) it is easy to show that the least-squares estimate of β_j is

$$\hat{\beta}_j = \frac{1}{n} \sum_{i=1}^n \varphi_j(x_i) Y_i, \quad j = 1, \dots, n.$$

If we assume the model like (4) and independent errors, then the coefficients $\hat{\beta}_1, \dots, \hat{\beta}_n$ are mutually independent and asymptotically

$$\hat{\beta}_j \sim N\left(\frac{1}{n} \sum_{i=1}^n \mu(x_i) \varphi_j(x_i), \frac{\sigma^2}{n}\right), \quad j = 1, \dots, n.$$

□

3 The MDL Model Selection and Averaging

In the MDL model selection we assume the approximating model (4) with normally distributed random errors, i.e. $\varepsilon \sim N_n(\mathbf{0}, \mathbf{I})$, where \mathbf{I} is the $n \times n$ identity matrix. The response data \mathbf{y} are modelled with the normal density functions

$$f(\mathbf{y}; \beta_m, \sigma_m^2) = \frac{1}{(2\pi\sigma_m^2)^{n/2}} \exp\left(-\frac{1}{2\sigma_m^2} \|\mathbf{y} - \mathbf{X}_m \beta_m\|^2\right), \quad (9)$$

where $\|\cdot\|$ denotes the Euclidean norm of a vector and $1 \leq m \leq M$. Under these assumptions $\hat{\beta}_m$ is the maximum likelihood (ML) estimate of β_m and

$$\hat{\sigma}_m^2 = \|\mathbf{y} - \hat{\mu}_m\|^2 / n$$

the ML estimate of σ^2 .

Consider the normalized maximum likelihood (NML) function (Rissanen 1996 and 2007, Barron, Rissanen and Yu 1998)

$$\hat{f}(\mathbf{y}; m) = \frac{f(\mathbf{y}; \hat{\theta}_m(\mathbf{y}))}{C(m)}, \quad (10)$$

where $\hat{\theta}_m = (\hat{\beta}_m', \hat{\sigma}_m^2)'$ and

$$C(m) = \int f(\mathbf{y}; \hat{\theta}_m(\mathbf{y})) d\mathbf{y} \quad (11)$$

is the normalizing constant. Thus $\hat{f}(\mathbf{y}; m)$ is a density function, provided that $C(m)$ is bounded. Rissanen (1996) considers the NML function in the context of coding and modelling theory and takes

$$-\log \hat{f}(\mathbf{y}; m) = -\log f(\mathbf{y}; \hat{\theta}_m(\mathbf{y})) + \log C(m) \quad (12)$$

as the “shortest code length” for the data \mathbf{y} that can be obtained with the model \mathcal{M}_m and calls it *the stochastic complexity* of \mathbf{y} , given \mathcal{M}_m . The last term in the equation (12) is called *the parametric complexity*.

Here we consider the model class

$$\mathcal{M}_m = \{f(\mathbf{y}; \theta_m) : m \in \{1, \dots, M\}\} \quad (13)$$

defined by the normal densities (9). The aim of variable selection is to find the optimal value of index m . According to *the MDL* (Minimum Description Length) *principle* we seek to find the index value $m = \hat{m}$ that minimizes the stochastic complexity:

$$-\log \hat{f}(\mathbf{y}; \hat{m}) = \min_m \{-\log f(\mathbf{y}; \hat{\boldsymbol{\theta}}_m(\mathbf{y})) + \log C(m)\}.$$

Since \hat{m} maximizes (10), we may call it the NML estimate of m within the model class \mathcal{M}_m .

For the normal distribution (9), however, the normalizing constant $C(m)$ is not bounded and hence the NML function is not defined. One approach to this problem is to constrain the data space properly (Rissanen 2000). For the constrained data space the stochastic complexity $C(m)$ is bounded, but it will depend on certain hyperparameters (Rissanen 2007 p. 116). The negative logarithm of $\hat{f}(\mathbf{y}; m)$ multiplied by 2 is given by

$$-2\log \hat{f}(\mathbf{y}; m) = n \log \hat{\sigma}_m^2 + k_m \log \frac{\|\hat{\boldsymbol{\mu}}_m\|^2}{\hat{\sigma}_m^2} - 2 \log \Gamma\left(\frac{n - k_m}{2}\right) - 2 \log \Gamma\left(\frac{k_m}{2}\right) + L(m) + c,$$

where the constant c is common to all models and therefore it can be ignored in model selection. The code length $L(m)$ for m is small and will be omitted. If we denote $\text{MDL}_m = -2 \log \hat{f}(\mathbf{y}; m)$ and omit $L(m) + c$, the NML model selection criterion takes the form (Hansen and Yu 2001, Liski 2006)

$$\text{MDL}_m = n \log S_m^2 + k_m \log F_m + \log[k_m(n - k_m)],$$

where $S_m^2 = \|\mathbf{y} - \hat{\boldsymbol{\mu}}_m\|^2 / (n - k_m)$ and $F_m = \|\hat{\boldsymbol{\mu}}_m\|^2 / (k_m S_m^2)$.

Consider a mixture density

$$\sum_{m=1}^M w_m \hat{f}(\mathbf{y}; m) \quad \text{with} \quad \sum_{m=1}^M w_m = 1,$$

where $\hat{f}(\mathbf{y}; m)$ are NML densities and w_m nonnegative weights $1 \leq m \leq M$. If we select the model $m = \hat{m}$ and encode the data using the selected model \hat{m} , then the code length for the data is $\log[1/w_{\hat{m}} \hat{f}(\mathbf{y}; \hat{m})]$. On the other hand, the mixture model yields the code length $\log[1/\sum_m w_m \hat{f}(\mathbf{y}; m)]$ which is always shorter if $w_{\hat{m}} \neq 1$. Therefore, it seems advantageous to encode with a mixture (cf. also Liang and Barron 2005). However, the problem of finding the weight vector still remains.

Given the data \mathbf{y} , $\hat{f}(\mathbf{y}; m)$ can be interpreted as the likelihood of the model \mathcal{M}_m , $m = 1, 2, \dots, M$. This leads to the NML distribution

$$\hat{p}(m; \mathbf{y}) = \frac{\hat{f}(\mathbf{y}; m)}{\sum_{i=1}^M \hat{f}(\mathbf{y}; i)} = \frac{\exp(-\text{MDL}_m/2)}{\sum_{i=1}^M \exp(-\text{MDL}_i/2)} \quad (14)$$

for models (13). Thus the MDL distribution (14) may be used to define the empirical selected weight vector

$$\hat{\mathbf{w}} = (\hat{p}(1; \mathbf{y}), \dots, \hat{p}(M; \mathbf{y}))' \quad (15)$$

(cf. Rissanen 2007, Subsection 5.2.2).

4 Alternative Model Average Estimators

It is well-known that a model selection procedure can be unstable, as small changes in the data may lead to significant changes in model choice. The inference done with a single estimate $\hat{\boldsymbol{\beta}}_m$

based on the chosen model \mathcal{M}_m does not take into account model uncertainty, and therefore may be too optimistic. To deal with uncertainty in model selection we study model average estimation. Let $\hat{\boldsymbol{\mu}}_w$ denote an MA estimator of $\boldsymbol{\mu}$. It is a convex combination of estimators (5) such that

$$\hat{\boldsymbol{\mu}}_w = \sum_{m=1}^M w_m \hat{\boldsymbol{\mu}}_m = \sum_{m=1}^M w_m \mathbf{H}_m \mathbf{y} = \left(\sum_{m=1}^M w_m \mathbf{H}_m \right) \mathbf{y} = \mathbf{H}_w \mathbf{y}, \quad (16)$$

where \mathbf{H}_w denotes the implied hat matrix $\sum_{m=1}^M w_m \mathbf{H}_m$. Note that although every hat matrix \mathbf{H}_m is idempotent, the implied hat matrix \mathbf{H}_w is generally not. Selecting the model weights by the NML distribution (15) yields an operational model average estimator.

Bayesian model averaging is widely used in the literature and so we refer to these works by, among others, Draper (1995) and for literature reviews see Hoeting, et al. (1999). An alternative can be based on the analogue of Bayesian model probabilities for frequentist statistics. Such a weigh scheme has been implied in a series of papers by Akaike (see e.g. Akaike 1978 and 1979) and expounded further by Buckland, et al. (1997) and Burnhan and Anderson (2002). Akaike's suggestion derives from the Akaike information criterion (AIC). The Akaike weights are defined as

$$w_m \propto \exp(-\text{AIC}_m / 2)$$

normalized to have unit sum. In the present context of ML estimation $\text{AIC}_m = n \log \hat{\sigma}_m^2 + 2k_m$. For a Bayesian the weights

$$w_m \propto \exp(-\text{BIC}_m / 2)$$

can serve as a rough approximation to the posterior probabilities for models \mathcal{M}_m , where $\text{BIC}_m = n \log \hat{\sigma}_m^2 + k_m \log n$ is the Bayesian information criterion (Schwarz 1978) for \mathcal{M}_m .

Hansen (2007) proposed the Mallows' criterion

$$C(\mathbf{w}) = \|\mathbf{y} - \hat{\boldsymbol{\mu}}_w\|^2 + 2\sigma^2 \mathbf{k}'\mathbf{w}, \quad (17)$$

where $\mathbf{k} = (k_1, \dots, k_M)'$. The empirical Mallows' weight vector $\hat{\mathbf{w}}$ is selected so that the criterion (17) attains its minimum. In practice σ^2 should be replaced with some consistent estimator. In the simulation experiments our choice is $\sigma^2 = \hat{\sigma}_M^2$. There is no closed form solution to minimizing of (17) and the weight vector must be found numerically.

Leung and Barron (2006) considered MA estimators under the Gaussian model (9) when σ^2 is assumed to be known. They defined the weights to be

$$w_m \propto \exp(-\alpha \frac{\hat{r}_m}{2\sigma^2}), \quad \alpha > 0, \quad (18)$$

where

$$\hat{r}_m = \|\mathbf{y} - \hat{\boldsymbol{\mu}}_m\|^2 + \sigma^2(2k_m - n) \quad (19)$$

is an unbiased estimate of the model error

$$r_m = \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_m\|^2 \quad (20)$$

in the sense that $E(\hat{r}_m) = E(r_m)$ (Akaike 1970 and 1973, Mallows 1973, Stein 1973 and 1981). The tuning parameter α adjusts the degree of concentration of the weights on the models with small model error estimates. They derived simple and accurate bounds on (20) and its estimate (19). The weights (18) are not directly operational, however, since σ^2 and α should be estimated. Note that also $C(\mathbf{w}) - n\sigma^2$ is an unbiased estimate of the model error in the sense that

$$E[C(\mathbf{w}) - n\sigma^2] = E(r_w),$$

where $r_w = \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_w\|^2$ (Hansen 2007).

5 Simulations

In this section we report on simulation investigations of the performance potential of the MDL, AIC, BIC and Mallows' MA estimators. The setting is the regression

$$y_i = \sum_{j=1}^K x_{ij}\beta_j + \varepsilon_i, \quad (21)$$

where $K = 100$ and $k_M < K$. We fix in (21) $x_{i1} = 1$, and the remaining elements x_{ij} are mutually independent and follow the normal distribution $N(0, 1)$. The errors ε_i are independent of x_{ij} and are $N(0, 1)$. We impose the structure of gradual decay on β and varied

$$\beta_j = c \frac{A(a)}{j^{a+1/2}}$$

with different values a and c . Here $A(a)$ is such a normalizing constant that $\text{Var}(\sum_{j=1}^K x_{ij}\beta_j) = c^2$. The parameter a controls the speed of decay and the coefficient c determines the value of $R^2 = c^2/(1 + c^2)$, a measure of explained variation. In the reported experiments the sample size is set to $n = 50$ and the maximum model order k_M varied between 10, 20, 30, 40, 45.

Let $\text{MDL}(\mathbf{w})$, $\text{AIC}(\mathbf{w})$, $\text{BIC}(\mathbf{w})$, $\text{MalC}(\mathbf{w})$ denote the MDL, AIC, BIC and Mallows' MA estimators, respectively. To evaluate estimators we compute the model error (20). We then summarise the overall performance by computing the average model error (AME) over 10 000 iterations in each of our various set-ups. We normalize the AME by that of the estimator $\hat{\beta}_M$, so that unity indicates equivalence with $\hat{\beta}_M$ in the AME sense. Then the AME curves as a function of R^2 are displayed.

In the first experiment (Figure 1) we illustrate the effect of a (the speed of decay) on the performance of estimators. The parameter a varied between 0.5, 1.0, 1.5 and 2 when $M = 40$ and $n = 50$. The results from the first experiment show that the performance of $\text{MDL}(\mathbf{w})$ and $\text{BIC}(\mathbf{w})$ are close to each others. Overall, the $\text{AIC}(\mathbf{w})$ estimator has clearly higher AME relative to its competitors. For large values of a (1.5 and 2) the $\text{MalC}(\mathbf{w})$ has slightly higher AME than $\text{MDL}(\mathbf{w})$ and $\text{BIC}(\mathbf{w})$, but for the values 0.5 and 1 there are some crossings of the AME curves. In all cases the AME curves are increasing functions of R^2 .

The second experiment (Figure 2) depicts the dependence of the AME on the maximal model order k_M that varied between 10, 20, 30 and 45. The main message is clear: the AME curves of the MA estimators $\text{MDL}(\mathbf{w})$, $\text{BIC}(\mathbf{w})$ and $\text{MalC}(\mathbf{w})$ are pretty close to each others and their performance relative to the $\text{AIC}(\mathbf{w})$ improves when k_M increases. Results of further simulation experiments (not reported here) confirmed the finding, that the performance of $\text{MDL}(\mathbf{w})$ and $\text{BIC}(\mathbf{w})$ relative to the $\text{AIC}(\mathbf{w})$ improves when M/n increases.

Note that although the $\text{AIC}(\mathbf{w})$ does not do too well in our experiments, its relative performance improves when M/n is small. Hansen (2007) reported simulation results showing that the $\text{AIC}(\mathbf{w})$ and $\text{MalC}(\mathbf{w})$ are superior to the $\text{BIC}(\mathbf{w})$ when M/n is small. In his experiments both M and n varied (K is sufficiently large).

Finally, in Figure 3 we illustrate by simulation how $\text{MDL}(\mathbf{w})$ is superior to the MDL model selection estimator in the AME sense. The $\text{MDL}(\mathbf{w})$ consistently outperforms the MDL model selection estimator.

References

Akaike, H. (1970). Statistical Predictor Identification. *Annals of the Institute of Statistical Mathematics*, 22, 203–217.

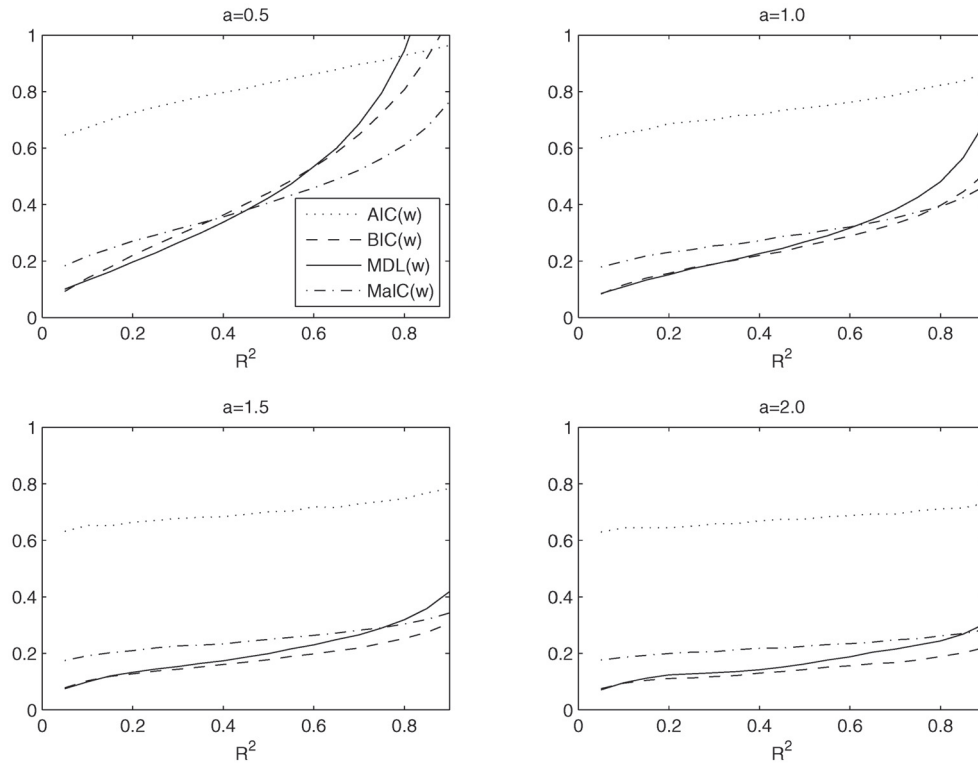


Figure 1: The AME curves of $MDL(w)$, $BIC(w)$, $AIC(w)$ and $MalC(w)$ as a function of R^2 for $a = 0.5, 1.0, 1.5$ and 2.0 when $n = 50$ and $M = 40$.

- Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. Pages 267–281 in B.N. Petrov, and F. Csaki, (eds.) *Second International Symposium on Information Theory*. Akademiai Kiado, Budapest.
- Akaike, H. (1978). A Bayesian Analysis of the Minimum AIC Procedure. *Annals of the Institute of Statistical Mathematics*, 30, 9–14.
- Akaike, H. (1979). A Bayesian Extension of the Minimum AIC Procedure to Autoregressive Model Fitting. *Biometrika*, 66, 237–242.
- Barron, A.R., Rissanen, J. and Yu, B. (1998). The MDL principle in modeling and coding. *Special Issue of Information Theory to Commemorate 50 Years of Information Theory*, 44, 2743–2760.
- Buckland, S.T. Burnham, K.P. and Augustin, N.H. (1999). Model Selection: An Integral Part of Inference. *Biometrics*, 53, 603–618.
- Burnham, K.P. and Anderson D.R. (2002). *Model Selection and Multi-model Inference*. New York, Springer-Verlag.
- Draper, D. (1995). Assessment and Propagation of Model Uncertainty. *Journal of the Royal Statistical Society, B* 57, 45–70.
- Efromovich, S. (1999). *Nonparametric Curve Estimation*. New York, Springer-Verlag.

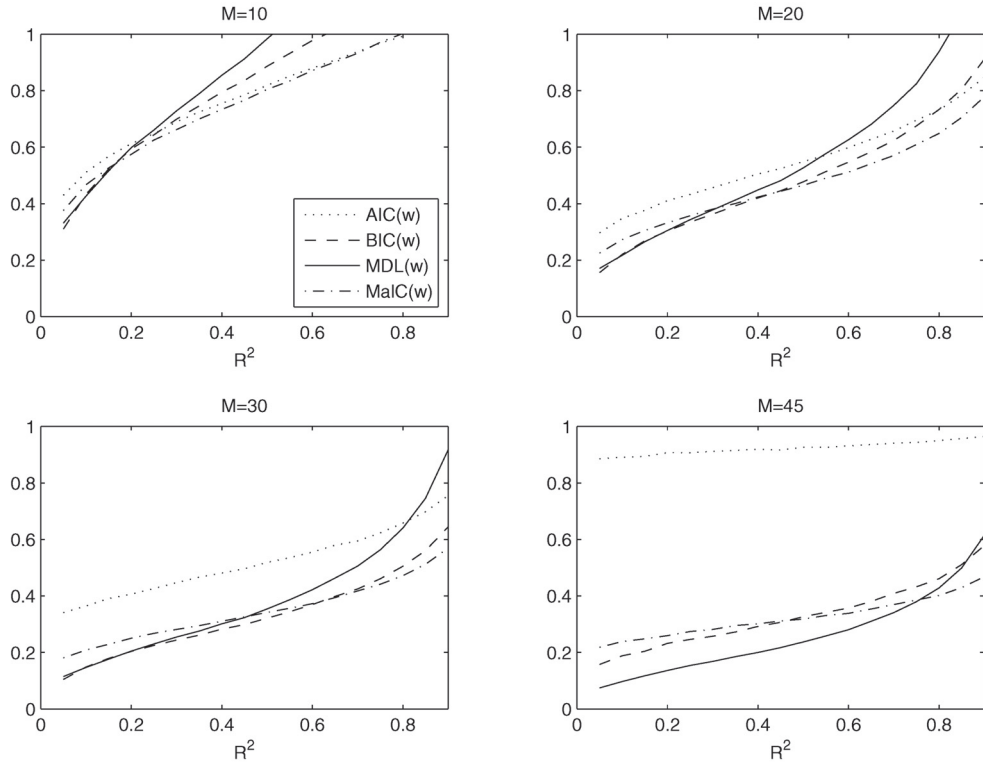


Figure 2: The AME curves of $\text{MDL}(w)$, $\text{BIC}(w)$, $\text{AIC}(w)$ and $\text{MalC}(w)$ as a function of R^2 for $M = 10, 20, 30$ and 45 when $n = 50$ and $a = 1$.

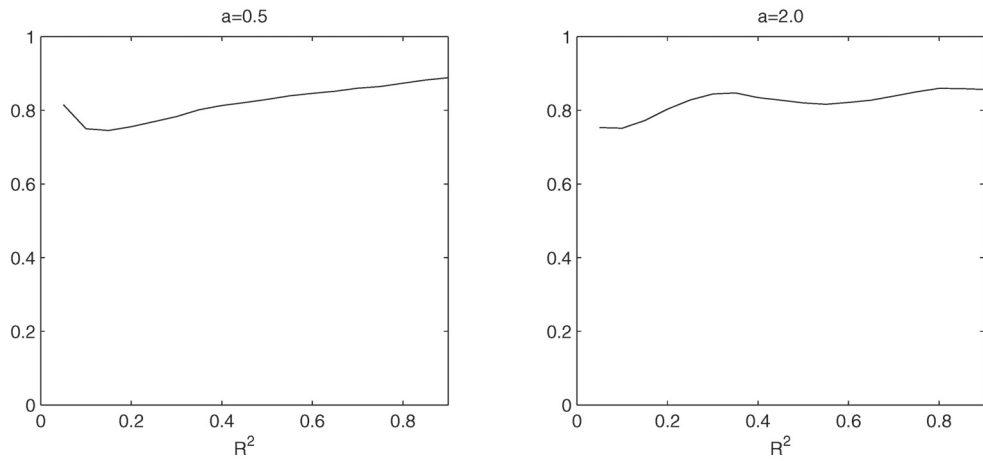


Figure 3: The AME of the $\text{MDL}(w)$ normalized by that of the MDL model selection estimator when $n = 50$ and $M = 40$.

- Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing*. New York, Springer-Verlag.
- Hansen, B. E. (2007). Least Squares Model Averaging. *Econometrica* 75 (4), 1175–1189.
- Hansen, A. J. and Yu, B. (2001). Model Selection and the Principle of Minimum Description Length. *Journal of the American Statistical Association*, 96, 746–774.
- Hjort, N. L. and Claeskens, G. (2003). Frequentist Model Average Estimators. *Journal of the American Statistical Association*, 98, 879–899.
- Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science*, 14, 382–417.
- Leung, G. and Barron, A. R. (2006). Information Theory and Mixing Least-Squares Regressions. *IEEE Transactions on Information Theory*, IT-52, No. 8, 3396–3410.
- Liang, F. and Barron, A. R. (2005) Exact Minimax Predictive Density Estimation and MDL. In: Grünwald, P. D., Myung, I. J. and Pitt, M. A. (Eds.). *Advances in Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT Press.
- Liski, E. P. (2006). Normalized ML and the MDL Principle for Variable Selection in Linear Regression In: Liski, E. P., Isotalo, J., Niemelä, J., Puntanen, S., and Styan, G. P. H. (Eds.). *Festschrift for Tarmo Pukkila on His 60th Birthday*, 159–172. Tampere, Department of Mathematics, Statistics and Philosophy.
- Mallows, C. L. (1973). Some comments on C_p . *Technometrics*, 15, 661–675.
- Newey, W. K. (1997). Convergence Rates and Asymptotic Normality for Series Estimators. *Journal of Econometrics*, 79, 147–168.
- Rissanen, J. (1978). Modeling by Shortest Data Description. *Automatica*, 14, No. 1, 465–471.
- Rissanen, J. (1986). Stochastic Complexity and Modeling. *Annals of Statistics*, 14, 1080–1100.
- Rissanen, J. (1996). Fisher Information and Stochastic Complexity. *IEEE Transactions on Information Theory*, IT-42, No. 1, 40–47.
- Rissanen, J. (2000). MDL Denoising. *IEEE Trans. on Information Theory*, IT-46, No. 1, 2537–2543.
- Rissanen, J. (2007). *Information and Complexity and in Statistical Modeling*. New York, Springer-Verlag.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.
- Stein, C. (1973). Estimation of the Mean of a Multivariate Normal Distribution. In: *Proceedings of the Prague Symposium in Asymptotic Statistics, 1973*, 345–381.
- Stein, C. (1981). Estimation of the Mean of a Multivariate Normal Distribution. *Annals of Statistics*, 9, 1135–1151.

[6] Liski, A. and Liski, E. P. (2013)

MDL model selection criterion for mixed models with an application to spline smoothing

In: SenGupta, A., Samanta, T. & Basu, A. (eds.) *Statistical Paradigms - Recent Advances and Reconciliations*, Statistical Science and Interdisciplinary Research - Vol 14, World Scientific, Accepted for publication.

Chapter 1

MDL model selection criterion for mixed models with an application to spline smoothing

Antti Liski and Erkki P. Liski*

*Tampere University of Technology, Department of Signal Processing,
P.O.Box 553, 33101 Tampere, FINLAND
antti.liski@tut.fi*

For spline smoothing one can rewrite the smooth estimation as a linear mixed model (LMM) where the smoothing parameter appears as the ratio between the variance of the error terms and the variance of random effects. Smoothing methods that use basis functions with penalization can utilize maximum likelihood (ML) theory in the LMM framework. We introduce the minimum description length (MDL) model selection criterion for LMM's and propose an automatic data-based spline smoothing method based on the MDL criterion. Simulation study shows that the performance of *MDL* in spline smoothing is close to that of the *BIC* criterion.

1.1. INTRODUCTION

This paper considers model selection for linear mixed models (LMM) using the MDL principle.¹⁻³ Regression splines that use basis functions with penalization can be fit conveniently using the machinery of LMMs, and thereby borrow from a rich source of existing methodology (cf. Refs. 4,5). The basis coefficients can be considered as random coefficients and the smoothing parameter as the ratio between variances of the error variables and random effects, respectively. In this article we present the MDL criterion under a LMM for choosing the number of knots, the amount of smoothing and the basis jointly. A simulation experiment was conducted to compare the performance of the MDL method with that of the corresponding techniques based on the Akaike information criterion *AIC*, corrected *AIC* (*AICc*), and generalized crossvalidation *GCV*.

*University of Tampere, Department of Mathematics and Statistics

The model known as the linear mixed model may be written as

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}, & \mathbf{b} &\sim \text{N}(\mathbf{0}, \phi^2 \mathbf{I}_m), \\ \boldsymbol{\varepsilon} &\sim \text{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n), & \text{Cov}(\mathbf{b}, \boldsymbol{\varepsilon}) &= \mathbf{0}, \end{aligned} \quad (1.1)$$

where \mathbf{X} and \mathbf{Z} are known $n \times p$ and $n \times m$ matrices, respectively, \mathbf{b} is the $m \times 1$ vector of random effects that occur in the $n \times 1$ data vector \mathbf{y} and $\boldsymbol{\beta}$ is the $p \times 1$ vector of unknown fixed effects parameters. Compared with the ordinary linear regression model, the difference is $\mathbf{Z}\mathbf{b}$, which may take various forms, thus creating a rich class of models. Then under these conditions we have

$$\mathbf{y} \sim \text{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{V}) \quad (1.2)$$

and

$$\mathbf{y}|\mathbf{b} \sim \text{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \sigma^2 \mathbf{I}_n), \quad (1.3)$$

where $\mathbf{V} = \frac{1}{\alpha} \mathbf{Z}\mathbf{Z}' + \mathbf{I}_n$ for $\alpha = \sigma^2/\phi^2 > 0$. The parameter α is the ratio between the variance of the error variables ε_i , $1 \leq i \leq n$ and the variance of the random effects b_j , $1 \leq j \leq m$. The set of possible values for α is $[0, \infty]$. There are different types of LMMs, and various ways of classifying them. For these we refer to large literature on mixed models (see e.g. Refs. 6,7).

The paper is organized as follows. In Section 1.2, we consider likelihood estimation in LMMs. In Subsection 1.2.2 the estimates of the fixed effects and random effects parameters are presented as a function of the smoothing parameter. The MDL model selection criterion is introduced in Section 1.3, and it is applied to automatic scatterplot smoothing in Section 1.4. Section 1.5 presents simulation results.

1.2. LIKELIHOOD ESTIMATION FOR LINEAR MIXED MODELS

In the LMM (1.1) the interest is either in the fixed effects parameter $\boldsymbol{\beta}$, or also in the associated random effects \mathbf{b} . If we focus only on the estimation of the vector of fixed effects $\boldsymbol{\beta}$, then we have the linear model (1.2) and the vector of random effects \mathbf{b} is a device for modelling the covariance structure for the response \mathbf{y} . In many applications, the random effects themselves are of interest. In this case the choice of fixed versus random effects is a legitimate modelling choice.

Let $h(\mathbf{b}; \sigma^2)$ denote the density function of the vector \mathbf{b} of random effects, and $f(\mathbf{y}|\mathbf{b}; \beta, \sigma^2)$ the conditional density function of \mathbf{y} given \mathbf{b} . Then the joint density function of \mathbf{y} and \mathbf{b} is

$$\begin{aligned} & f(\mathbf{y}|\mathbf{b}; \beta, \sigma^2)h(\mathbf{b}; \sigma^2) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{b}\|^2\right) \\ & \quad \times \left(\frac{\alpha}{2\pi\sigma^2}\right)^{m/2} \exp\left(-\frac{\alpha}{2\sigma^2}\|\mathbf{b}\|^2\right) \\ &= \frac{\alpha^{m/2}}{(2\pi\sigma^2)^{(n+m)/2}} \exp\left[-\frac{1}{2\sigma^2}(\|\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{b}\|^2 + \alpha\|\mathbf{b}\|^2)\right]. \end{aligned} \quad (1.4)$$

The likelihood function for the model (1.1) is the density function (1.4) viewed as a function of the parameters β and σ^2 for fixed data \mathbf{y} . Since the nonobservable vector \mathbf{b} of random effects is part of the model, we integrate the joint density (1.4) with respect to \mathbf{b} . The function

$$L(\beta, \sigma^2; \mathbf{y}) = \int f(\mathbf{y}|\mathbf{b}; \beta, \sigma^2)h(\mathbf{b}; \sigma^2) d\mathbf{b} \quad (1.5)$$

is the integrated likelihood function corresponding to the normal density $h(\mathbf{b}; \sigma^2)$. The likelihood (1.5) takes the form ⁸

$$L(\beta, \sigma^2; \mathbf{y}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}(\|\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\tilde{\mathbf{b}}\|^2 + \alpha\|\tilde{\mathbf{b}}\|^2)\right] |\mathbf{V}|^{-1/2}, \quad (1.6)$$

where $\tilde{\mathbf{b}} = (\mathbf{Z}'\mathbf{Z} + \alpha\mathbf{I}_m)^{-1}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\beta)$.

The vector denoted by $\tilde{\mathbf{b}}$ in the function (1.6) can be thought of as a parameter vector just as β . The likelihood function (1.6) is used to determine the ML estimates of β and σ^2 as well as to estimate $\tilde{\mathbf{b}}$. Twice the logarithm of the likelihood function (1.6) is

$$2 \log[L(\beta, \sigma^2)] = -n \log(\sigma^2) - \frac{1}{\sigma^2} \|\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\tilde{\mathbf{b}}\|^2 - \alpha\|\tilde{\mathbf{b}}\|^2, \quad (1.7)$$

where the unnecessary constants are omitted.

1.2.1. Mixed model equations

The function (1.7) can be considered as a penalized log-likelihood function. For a given α , the penalized maximum likelihood estimators for β and $\tilde{\mathbf{b}}$ from (1.7) are equivalent to the solution of the so-called mixed model equations (e.g. Ref. 7)

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \alpha\mathbf{I}_m \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{\mathbf{b}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{pmatrix}, \quad (1.8)$$

which yield the estimates

$$\hat{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}, \quad (1.9)$$

$$\hat{\mathbf{b}} = (\mathbf{Z}'\mathbf{Z} + \alpha\mathbf{I}_m)^{-1}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\hat{\beta}). \quad (1.10)$$

The mixed model equations (1.8) refer to the LMM (1.1) which is an extension of the ordinary regression model. In Ref. 9 it was shown that the derived estimates are indeed the best linear unbiased predictors (BLUP). In Ref. 10 a wide ranging account of mixed model equations and BLUP are given with examples, applications and discussion.

Let $\delta = (\beta', \mathbf{b}')'$, $\mathbf{M} = (\mathbf{X}, \mathbf{Z})$ and $\mathbf{D} = \text{diag}(0, \dots, 0, 1, \dots, 1)$ a $(p + m) \times (p + m)$ diagonal matrix, whose first p diagonal elements are zero and the other m diagonal elements are 1. Then by (1.8)

$$\hat{\delta} = \begin{pmatrix} \hat{\beta} \\ \hat{\mathbf{b}} \end{pmatrix} = (\mathbf{M}'\mathbf{M} + \alpha\mathbf{D})^{-1}\mathbf{M}'\mathbf{y} \quad (1.11)$$

and the ordinary least squares estimate

$$\tilde{\delta} = (\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'\mathbf{y}$$

of δ is obtained by putting $\alpha = 0$. Hence, for a given α , $\hat{\delta}$ is the linear transformation

$$\hat{\delta} = \mathbf{B}\tilde{\delta} \quad (1.12)$$

of $\tilde{\delta}$, where $\mathbf{B} = (\mathbf{M}'\mathbf{M} + \alpha\mathbf{D})^{-1}\mathbf{M}'\mathbf{M}$ is a shrinkage matrix whose eigenvalues lie in $[0, 1]$. Thus (1.12) is a ridge type estimator. Under the model (1.3)

$$\hat{\delta} \sim \mathbf{N}[\mathbf{B}\delta, \sigma^2\mathbf{B}(\mathbf{M}'\mathbf{M})^{-1}\mathbf{B}]. \quad (1.13)$$

Maximizing the log-likelihood (1.7) with respect to σ^2 and inserting the estimators (1.9) and (1.10) for β and \mathbf{b} provide the estimate

$$\begin{aligned} \hat{\sigma}^2 &= n^{-1}\|\mathbf{y} - \mathbf{X}\hat{\beta} - \mathbf{Z}\hat{\mathbf{b}}\|^2 = n^{-1}\|\mathbf{y} - \hat{\mathbf{y}}\|^2 \\ &= n^{-1}\mathbf{y}'(\mathbf{I} - \mathbf{H})^2\mathbf{y}, \end{aligned} \quad (1.14)$$

where the fitted values are

$$\hat{\mathbf{y}} = \mathbf{M}\hat{\delta} = \mathbf{H}\mathbf{y} \quad (1.15)$$

and the hat matrix \mathbf{H} can be written as

$$\mathbf{H} = \mathbf{M}(\mathbf{M}'\mathbf{M} + \alpha\mathbf{D})^{-1}\mathbf{M}'. \quad (1.16)$$

Unlike for an ordinary linear regression model, \mathbf{H} is not a projection matrix for $\alpha > 0$.

The conditional distribution of $\mathbf{y}|\mathbf{b}$ corresponding to (1.3) yields the normal density function

$$f(\mathbf{y}|\mathbf{b}; \boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{M}\boldsymbol{\delta}\|^2\right). \quad (1.17)$$

Here \mathbf{b} is considered as a parameter vector just as $\boldsymbol{\beta}$. The estimators for $\boldsymbol{\delta}$ and σ^2 are given by (1.11) and (1.14), respectively.

1.2.2. Profile likelihood estimates

Note that $\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}}$ and $\hat{\sigma}^2$ are profile likelihood estimates depending on the value of α . The inverse of \mathbf{V} can be written as follows

$$\mathbf{V}^{-1} = (\mathbf{I}_n + \frac{1}{\alpha} \mathbf{Z}\mathbf{Z}')^{-1} = \mathbf{I}_n - \mathbf{Z}(\alpha\mathbf{I}_m + \mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'. \quad (1.18)$$

If $\alpha \rightarrow 0$, then $\mathbf{V}^{-1} \rightarrow \mathbf{I}_n - \mathbf{Z}\mathbf{Z}^+$, where $\mathbf{Z}^+ = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ is the Moore-Penrose inverse of \mathbf{Z} . Then using the above result, we conclude that $\hat{\boldsymbol{\beta}}$ approaches to

$$\hat{\boldsymbol{\beta}}_0 = [\mathbf{X}'(\mathbf{I}_n - \mathbf{Z}\mathbf{Z}^+)\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{I}_n - \mathbf{Z}\mathbf{Z}^+)\mathbf{y}, \quad (1.19)$$

as $\alpha \rightarrow 0$. Similarly, it follows from (1.10) that

$$\hat{\mathbf{b}} \rightarrow \hat{\mathbf{b}}_0 = \mathbf{Z}^+(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_0) \quad \text{as } \alpha \rightarrow 0. \quad (1.20)$$

Using (1.18), we have for $\mathbf{Z}'\mathbf{V}^{-1}$ the formula

$$\begin{aligned} \mathbf{Z}'(\mathbf{I}_n + \frac{1}{\alpha} \mathbf{Z}\mathbf{Z}')^{-1} &= \mathbf{Z}' - \mathbf{Z}'\mathbf{Z}(\alpha\mathbf{I}_m + \mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' \\ &= (\alpha\mathbf{I}_m + \mathbf{Z}'\mathbf{Z} - \mathbf{Z}'\mathbf{Z})(\alpha\mathbf{I}_m + \mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' \\ &= \alpha(\alpha\mathbf{I}_m + \mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}', \end{aligned} \quad (1.21)$$

which together with (1.10) implies

$$\hat{\mathbf{b}} = \frac{1}{\alpha} \mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \quad (1.22)$$

Consequently, (1.22) is the conditional expectation $E(\mathbf{b}|\mathbf{y})$ where $\boldsymbol{\beta}$ is replaced with $\hat{\boldsymbol{\beta}}$. Hence $\hat{\mathbf{b}} = \widehat{E(\mathbf{b}|\mathbf{y})}$ is the ML estimate of the mean of \mathbf{b} given a set of observations \mathbf{y} . If $\alpha \rightarrow \infty$, then clearly $\mathbf{V} \rightarrow \mathbf{I}_n$ and $\hat{\mathbf{b}} \rightarrow \mathbf{0}$. Thus clearly $\hat{\boldsymbol{\beta}}$ approaches to the ordinary least squares estimator

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad \text{as } \alpha \rightarrow \infty. \quad (1.23)$$

1.3. MODEL SELECTION IN LINEAR MIXED MODELS USING MDL CRITERION

1.3.1. Model selection

There is often uncertainty about which explanatory variables to use in \mathbf{X} , or how to select the matrix \mathbf{Z} . Typically we have a set of candidate models and the problem of model selection arises when one wants to decide which model to choose.

Let the variable η index the set of candidate models. We consider a set of conditional normal models corresponding to (1.3):

$$\mathbf{y}|\mathbf{b}_\eta \sim N(\mathbf{X}_\eta\boldsymbol{\beta}_\eta + \mathbf{Z}_\eta\mathbf{b}_\eta, \sigma^2\mathbf{I}_n), \quad (1.24)$$

where \mathbf{X}_η and \mathbf{Z}_η are $n \times p_\eta$ and $n \times m_\eta$ matrices, respectively, corresponding to the candidate model η . Here $\boldsymbol{\beta}_\eta$ and \mathbf{b}_η are $n \times p_\eta$ and $n \times m_\eta$ parameter vectors for the model η . Note that the estimates $\hat{\boldsymbol{\beta}}_\eta$, $\hat{\mathbf{b}}_\eta$ and $\hat{\sigma}^2$ depend on the tuning parameter $\alpha \in [0, \infty]$. In this conditional framework we specify a model by giving the pair (η, α) and denote $\gamma = (\eta, \alpha)$.

1.3.2. Normalized maximum likelihood

In Ref. 1 Rissanen developed an MDL criterion based on the normalized maximum likelihood (NML) coding scheme (cf. Ref. 11). Assume that the response data are modelled with a set of density functions $f(\mathbf{y}; \gamma, \boldsymbol{\theta})$, where the parameter vector $\boldsymbol{\theta}$ varies within a specified parameter space. The NML function is defined by

$$\hat{f}(\mathbf{y}; \gamma) = \frac{f(\mathbf{y}; \gamma, \hat{\boldsymbol{\theta}})}{C(\gamma)}, \quad (1.25)$$

where $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{y})$ is the ML estimator of $\boldsymbol{\theta}$ and

$$C(\gamma) = \int f(\mathbf{x}; \gamma, \hat{\boldsymbol{\theta}}(\mathbf{x})) d\mathbf{x} \quad (1.26)$$

is the normalizing constant. The integral in (1.26) is taken over the sample space. Thus $\hat{f}(\mathbf{y}; \gamma)$ defines a density function, provided that $C(\gamma)$ is bounded.

The expression

$$-\log \hat{f}(\mathbf{y}; \gamma) = -\log f(\mathbf{y}; \gamma, \hat{\boldsymbol{\theta}}) + \log C(\gamma) \quad (1.27)$$

is taken as the "shortest code length" for the data \mathbf{y} that can be obtained with the model γ and it is called *the stochastic complexity* of \mathbf{y} , given γ .¹

Here the estimate $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}}, \hat{\sigma}^2)$ is given by (1.9), (1.10) and (1.14). The last term in the equation (1.27) is called *the parametric complexity*.

It is clear that the NML function (1.25) attains its maximum and the "code length" (1.27) its minimum for the same value of γ . According to the MDL principle we seek the value $\gamma = \hat{\gamma}$ that minimizes the stochastic complexity (1.27). In general, obtaining the $\hat{\gamma}$ may be computationally a very intensive task.

Here the NML density (1.25) is needed for the model (1.24). However, the normalizing constant (1.26) for the model (1.24) is not finite. Following Rissanen's renormalizing approach^{2,3}, data \mathbf{y} is restricted to lie within a subset

$$\mathcal{Y}(s, R) = \{\mathbf{y} : \hat{\sigma}^2 \geq s, \hat{\boldsymbol{\delta}}' \mathbf{M}' \mathbf{M} \hat{\boldsymbol{\delta}} \leq nR\}, \quad (1.28)$$

where $s > 0$ and $R > 0$ are hyperparameters. Under the restriction (1.28) we have the NML density function

$$\hat{f}(\mathbf{y}; \gamma, s, R) = f(\mathbf{y}; \gamma, \hat{\boldsymbol{\theta}}) / C(s, R), \quad (1.29)$$

where $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\delta}}, \hat{\sigma}^2)$. For the model (1.24) the numerator in (1.29) takes a simple form

$$f(\mathbf{y}; \gamma, \hat{\boldsymbol{\theta}}) = (2\pi\hat{\sigma}^2 e)^{-\frac{n}{2}},$$

but the normalizing constant $C(s, R)$ will essentially depend on two hyperparameters s and R .

The code length (1.27) corresponding to (1.29) is minimized by setting $s = \hat{s} = \hat{\sigma}^2$ and $R = \hat{R} = \hat{\boldsymbol{\delta}}' \mathbf{M}' \mathbf{M} \hat{\boldsymbol{\delta}} / n$, i.e. by maximizing the NML density (1.29) with respect to s and R under the restriction (1.28). The explicit formula of $C(s, R)$ is given in the Appendix (formula (1.39)). Since $\hat{f}(\mathbf{y}; \gamma, \hat{\sigma}^2(\mathbf{y}), \hat{R}(\mathbf{y}))$ of (1.29) is not a density function, we normalize it. To keep the normalizing constant finite, the sample space is restricted such that $\hat{\sigma}^2 \in [s_1, s_2]$ and $\hat{R} \in [R_1, R_2]$, where $0 < s_1 < s_2$ and $0 < R_1 < R_2$ are hyperparameters. The resulting NML function

$$\hat{f}(\mathbf{y}; \gamma) = \hat{f}(\mathbf{y}; \gamma, \hat{\sigma}^2(\mathbf{y}), \hat{R}(\mathbf{y})) / C(\gamma),$$

is a density function, where the normalizing constant $C(\gamma)$ depends on the hyperparameters. Although the codelength will again depend on hyperparameters, they do not have essential effect on model selection. Derivation of the NML function for (1.24) under the LMM resembles that of the ordinary Gaussian linear regression.^{2,3}

1.3.3. MDL criterion

We are seeking models γ that minimize the "code length" $\log[1/\hat{f}(\mathbf{y}; \gamma)] = -\log \hat{f}(\mathbf{y}; \gamma)$. So, we define the selection criterion as $MDL(\gamma) = -2 \log \hat{f}(\mathbf{y}; \gamma)$, where the multiplier 2 is chosen just for convenience. For the model (1.24) under the *LMM* the *MDL* selection criterion takes the form

$$MDL(\gamma) = (n - d) \log \hat{\sigma}^2 + d \log \hat{R} - 2 \log \Gamma\left(\frac{n - d}{2}\right) - 2 \log \Gamma\left(\frac{d}{2}\right) \quad (1.30)$$

where $d = \text{tr } \mathbf{H}$ defines the model's degrees of freedom and $\hat{R} = \|\hat{\mathbf{y}}\|^2/n$. Note that $p \leq d \leq p + m$, $d = p$, as $\alpha = 0$ and $d \rightarrow p + m$, as $\alpha \rightarrow \infty$.

If we apply Stirling's approximation

$$\Gamma(x + 1) \approx (2\pi)^{1/2} (x + 1)^{x+1/2} e^{-x-1}$$

to the Γ -functions in (1.30) and omit the unnecessary constants, the criterion (1.30) takes the form

$$MDL(\gamma) = (n - d) \log \frac{\hat{\sigma}^2}{n - d} + d \log \frac{\hat{R}}{d} + \log[d(n - d)].$$

The derivation of the criterion (1.30) is outlined in the Appendix. In the extreme cases $\alpha = 0$ and $\alpha \rightarrow \infty$, the criterion (1.30) reduces to the ordinary Gaussian regression with $p + m$ and p regressors, respectively.

1.4. SPLINE SMOOTHING USING MDL CRITERION

Suppose the smoothing model

$$y_i = r(x_i) + \sigma \varepsilon_i, \quad i = 1, \dots, n, \quad (1.31)$$

where y_i is the observation for the i th subject, x_i is a scalar covariate, $r(\cdot)$ is a smooth function giving the conditional mean of y_i given x_i and $\varepsilon_1, \dots, \varepsilon_n$ are independent normally distributed error terms, i.e. $\varepsilon_i \sim N(0, 1)$. To pursue estimation, $r(\cdot)$ is replaced by a parametric regression spline model

$$r(x; \boldsymbol{\beta}, \mathbf{b}) = \beta_1 + \beta_2 x + \dots + \beta_p x^{p-1} + \sum_{j=1}^m b_j z_j(x). \quad (1.32)$$

The first p terms are a $(p - 1)$ th order polynomial of x , covariates $z_1(x), \dots, z_m(x)$ are elements of a smoothing basis, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$

and $\mathbf{b} = (b_1, \dots, b_m)'$ are unknown parameters. Then (1.32) can be written as

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \mathbf{z}_i' \mathbf{b} + \sigma \varepsilon_i,$$

where $\mathbf{x}_i = (1, x_i, \dots, x_i^{p-1})'$ and $\mathbf{z}_i = (z_1(x_i), \dots, z_m(x_i))'$. Typically \mathbf{x}_i is low-dimensional and \mathbf{z}_i is high-dimensional basis linearly independent of \mathbf{x}_i . A convenient choice is to use the truncated power basis of degree $p-1$. Then the i th row of \mathbf{Z} is $\mathbf{z}_i = ((x_i - \kappa_1)_+^{p-1}, \dots, (x_i - \kappa_m)_+^{p-1})$ with x_+ as positive part, so that for any number x , x_+ is x if x is positive and is equal to 0 otherwise. The knots $\kappa_1, \dots, \kappa_m$ are fixed values covering the range of x_1, \dots, x_n .

The amount of smoothing is controlled by α , which is here referred to as a smoothing parameter. The fitted values for a spline regression are given by (1.15). In addition to the value of α , the degree of the regression spline and the number and location of knots must be specified. Here we adopt the procedure where the knots are located at "equally spaced" sample quantiles of x_1, \dots, x_n . Thus the k th knot is the j th order statistic of $x_{(1)}, \dots, x_{(n)}$ where j is $nk/(m+1)$ rounded to the nearest integer. As soon as the degree of the regression spline is specified, one has to fix the number of knots. It is often recommended to choose basis in a "generous" manner such that there are enough knots to fit features in the data (see e.g. Ref. 12). The relation between spline smoothing and mixed models in general has been discussed in Ref. 13, for example. Penalized spline estimation for smoothing was made popular in statistics by Eilers and Marx.¹⁴

In smoothing we control three modeling parameters: the degree of the regression spline $p-1$, the number of knots m and the smoothing parameter α . A model $\gamma = (p, m, \alpha)$ is specified by the triple where the values for the modeling parameters p, m and α should be determined in an optimal way. The choice of α has a profound influence on the fit. In fact, it was shown in Subsection 1.2.2 that α can be chosen to give any one of a spectrum of fits between the unconstrained regression spline and the least-squares polynomial fit. As $\alpha \rightarrow \infty$, the regression spline approach by (1.23) to a smooth polynomial. The case $\alpha = 0$ corresponds to the unconstrained case where the estimates of $\boldsymbol{\beta}$ and \mathbf{b} are given by (1.19) and (1.20), respectively. A model estimator $\hat{\gamma}$ is obtained by minimizing the the MDL selection criterion (1.30) with respect to model $\gamma = (p, m, \alpha)$, that is, with respect to parameters p, m and α , using numerical optimization routines.

1.5. SIMULATIONS

1.5.1. Preamble

In this section we give an outline of a simulation study which aims at the comparison of the performance of several model selection techniques in data based smoothing. Apart from smoothing, the number of knots is specified automatically. Along with the *MDL* criterion, we briefly review the performance of the model selection criteria *AICc* (corrected *AIC*), *BIC* (Bayesian information criterion) and *GCV* (generalized cross-validation, see Ref. 5).

In all investigated scenarios, we considered the model given in (1.32), where the x_i , $i = 1, \dots, n$, were equally spaced on $[0, 1]$. Four different regression functions $r(\cdot)$ were studied: the first, called "Logit", uses a logistic function

$$r(x) = 1/\{1 + \exp[-20(x - 0.5)]\}$$

and the second function "Bump" was

$$r(x) = x + 2 \exp\{-16(x - 0.5)^2\}.$$

The third function "SpaHetj" is

$$r(x) = \sqrt{x(1-x)} \sin\left(\frac{2\pi(1 + 2^{(9-4j)/5})}{x + 2^{(9-4j)/5}}\right),$$

where the parameter $j = 1, 2, \dots$ controls spatial variation. The value $j = 1$ (SpaHet1) yields low spatial variation and larger values of j (eg. SpaHet3) imply greater spatial heterogeneity. The fourth function "Sinj"

$$r(x) = \sin(2\pi\theta), \quad \theta = j$$

is a cyclic function, where the parameter θ controls the number of cycles. Ruppert¹² used the above mentioned functions, among all, in his simulation study.

The knots were located at equally spaced sample quantiles, so that the number of knots determines the knot sequence. In this study, only the first degree ($p - 1 = 1$), quadratic ($p - 1 = 2$) and cubic splines ($p - 1 = 3$) were considered. A model is specified by the triple $\gamma = (p, m, \alpha)$. For each combination of p and m the selection criterion was minimized with respect to α to determine the optimal model $\hat{\gamma}$. For each setting 500 datasets were simulated. The performance of the criteria were assessed by using the

function $MSE(x)$ defined as the mean over the generated datasets of the squared error

$$SE(x; \hat{\gamma}) = [r(x; \hat{\gamma}) - r(x)]^2 \quad (1.33)$$

at the point x , $MASE$ defined as

$$MASE = \sum_{i=1}^n MSE(x_i)/n, \quad (1.34)$$

and the average squared error

$$ASE(\hat{\gamma}) = \sum_{i=1}^n SE(x_i; \hat{\gamma}) \quad (1.35)$$

of a model $\hat{\gamma}$ for a given dataset.

Along with the MDL criterion, also the criteria $AICc$, BIC and GCV were used to choose an appropriate spline smoothing model (see Ref. 5) and the performance of these four criteria were compared. The corrected AIC criterion proposed in Ref. 15 is given by

$$AICc(\gamma) = \log RSS(\gamma) + \frac{2[d(\gamma) + 1]}{n - d(\gamma) - 2},$$

where $d(\gamma) = tr\mathbf{H}(\gamma)$ and $RSS(\gamma)$ is the residual sum of squares. The criterion known as generalized cross-validation (GCV) is

$$GCV(\gamma) = \log RSS(\gamma)/[1 - d(\gamma)/n]^2.$$

The Bayesian information criterion (BIC) is given by

$$BIC(\gamma) = \log RSS(\gamma) + \frac{d(\gamma) \log n}{n}.$$

The model selection criteria are minimized numerically and the model $\hat{\gamma}$ that minimizes the criterion is selected. Bump, Logit, Sin3 and Spa-Het3 functions were estimated using the spline model (1.32). An appropriate smoothing model was selected by using $AICc$, GCV , BIC and MDL criteria respectively, and the models were fitted to all simulated datasets. The performance of GCV is very close to that of $AICc$, but $AICc$ was uniformly better than GCV with respect to the $MASE$ criterion (1.34). Therefore the results for GCV are not reported in this paper.

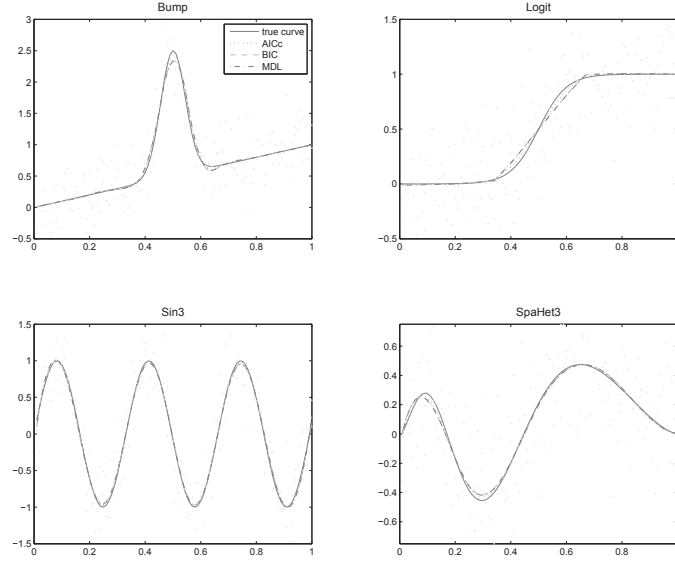


Fig. 1.1. The average fitted curves over 500 simulated data sets for the selection criteria $AICc$, BIC and MDL when the sample size $n = 200$.

1.5.2. Results

Inspection of Figure 1 shows that the average curves follow the true curves quite closely. However, in panel 2 each average curve tends to "straighten" the Logit function. It is obvious that the selected number of knots is not enough for the S-part of Logit function. Panel 1 shows that all criteria underestimate the bump part of the function, but underestimation is clearly greater when using MDL and BIC . These two criteria also react slower when recovering from the bump at 0.6 in Panel 1. In panel 4 all criteria slightly underestimate the changes in SpaHet3 function.

In Table 1 the $MASE$ values are reported for $AICc$, BIC and MDL under various settings. The degree of the fitted spline model chosen by the four criteria varies from one to three. The degree reported in Table 1 is the most frequently selected one under a given setting. When computing the value of $MASE$ for a given criterion, say MDL , the model

<i>function</i>	<i>n</i>	<i>degree</i>	<i>AICc</i>	<i>BIC</i>	<i>MDL</i>
Bump	50	1	0.0284(1.33)	0.0303(1.39)	0.0319(1.38)
	100	1	0.0133(1.39)	0.0135(1.37)	0.0141(1.41)
	200	2	0.0078(1.28)	0.0082(1.28)	0.0083(1.29)
Logit	50	1	0.0116(1.65)	0.0115(1.62)	0.0119(1.65)
	100	1	0.0066(1.73)	0.0057(1.41)	0.0057(1.41)
	200	1	0.0036(1.67)	0.0033(1.40)	0.0033(1.41)
Sin3	50	2	0.0224(1.33)	0.0211(1.24)	0.0224(1.32)
	100	2	0.0116(1.32)	0.0107(1.19)	0.0109(1.21)
	200	3	0.0064(1.29)	0.0073(1.49)	0.0072(1.47)
Spahet3	50	3	0.0154(1.40)	0.0170(1.55)	0.0153(1.39)
	100	3	0.0079(1.38)	0.0083(1.43)	0.0078(1.34)
	200	3	0.0040(1.41)	0.0038(1.32)	0.0038(1.32)

$\hat{\gamma} = (\hat{p}, \hat{m}, \hat{\alpha})$ for each data set is determined by minimizing the *MDL* criterion. Then *MASE* is obtained as the average over the $ASE(\hat{\gamma})$ values. Besides *MASE*, also relative *MASE* is reported. For computing the relative *MASE*, the minimum of the function $ASE(m) = ASE(\hat{p}, m, \hat{\alpha})$, say ASE^* , is determined with respect to the number of knots m for each data set. $MASE^*$ denotes the average of the ASE^* values over the generated datasets. Relative *MASE* is defined as the ratio $MASE/MASE^*$. Clearly $ASE^* \leq ASE(\hat{\gamma})$ for each $\hat{\gamma}$, and consequently relative *MASE* is not less than 1.

Inspection of the results in Table 1 shows that on the average the performance of *MDL* and *BIC* gets closer to each other as the sample size grows. This trend continues if we keep increasing the sample size n over 200. *BIC* and *MDL* do better than *AICc* for Logit and Spahet3 with $n = 200$. A large value of relative *MASE* indicates that the value of *MASE* can be considerably decreased by choosing the number of knots optimally. In view of the relative *MASE*, *BIC* and *MDL* are closer to optimal knot selection than *AICc*, except in case of Bump and Sin3 (when $n = 200$). Most of the time *BIC* and *MDL* yield relative *MASE*'s very close to each other.

Figure 2 displays histograms of the values of m chosen by the criteria *AICc*, *MDL*, *BIC* and *ASE* for SpaHet3 (500 datasets are generated). The *ASE* criterion uses the "oracle estimator" $\tilde{\gamma} = (\tilde{p}, \tilde{m}, \tilde{\alpha})$ that minimizes $ASE(\gamma)$. *ASE* chooses $m = 3$ in the vast majority of datasets. The behavior of *BIC* and *MDL* is closest to that of *ASE*. *AICc* tends to choose larger values of m than *BIC* and *MDL*. The corresponding behavior remains also when data are generated from Logit and Bump (not reported here). All model selection criteria tend to choose less knots than *ASE* when

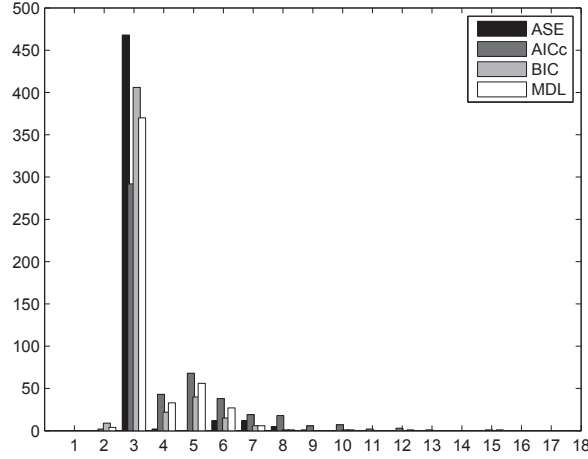


Fig. 1.2. Histograms of m as chosen by ASE , $AICc$, BIC and MDL criteria - *Spahet3*, $n=200$.

data are generated from Sin3 function (Figure 3). BIC and MDL tend to choose even less knots than $AICc$. One can see that the knot selection behavior of BIC is close to that of MDL .

In Figure 4 the graphs of the function $MSE(x)$ are displayed for all criteria and functions under consideration. Again we observe that MDL and BIC are very close to each other. It is also evident that $AICc$ tends to react to function fluctuations more aggressively than BIC and MDL . MDL and BIC seem to need more observations than $AICc$ to detect sudden changes in a function. We may note that the absence of outliers seems to give some advantage to $AICc$ and GCV over BIC and MDL .

In Figure 5 the ASE values (1.35) of MDL are plotted against the ASE values of $AICc$ and BIC , respectively, when datasets are generated from SpaHet3 (panels 1 and 2) and Sin3 (panels 3 and 4). In Panel 1 most of the ASE values are concentrated near the 45° line but MDL did clearly better than $AICc$ more often ($r = 0.89$). In the scatterplot MDL versus BIC (panel 2) the values are nicely concentrated on the 45 degree line, except a couple of outliers ($r = 0.93$). In Panel 3 the majority of the ASE values lie on the upper side of the 45 degree line ($r = 0.81$) and the scatterplot in panel 4 again refers to the similar performance of MDL and

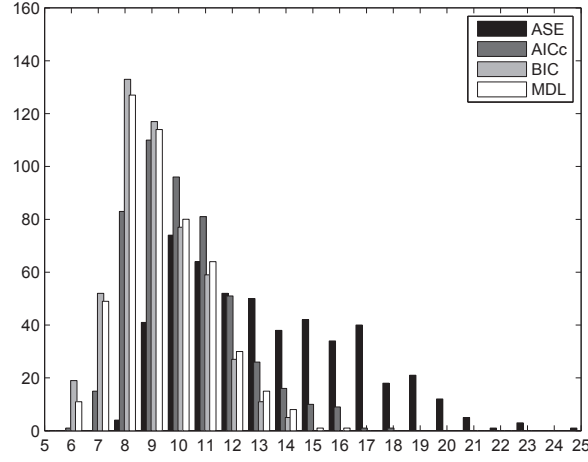


Fig. 1.3. Histograms of m as chosen by ASE , $AICc$, BIC and MDL criteria - $Sin3$, $n=200$.

BIC ($r = 0.96$). The ASE values lie very close to the 45 degree line except 11 cases where BIC fails. These outliers have an effect on the $MASE$ value as can be seen from Table 1.

1.6. CONCLUSIONS

We have derived the MDL model selection criterion in the context of linear mixed models. It is an extension of the corresponding criterion known in linear regression. Spline smoothing is formulated as an estimation problem within the context of linear mixed models. Then an automatic MDL procedure for choosing the smoothing parameter, the number of knots and the smoothing basis is presented as a model selection problem. The performance of MDL is compared with the $AICc$, BIC and GCV criteria. The simulation studies show that the results between the MDL approach and other methods are comparable in all cases. Furthermore, the performance of MDL is very close to that of BIC . No criterion dominates the other criteria uniformly. The MDL procedure outperforms the other methods in the case of $SpaHet3$ function.

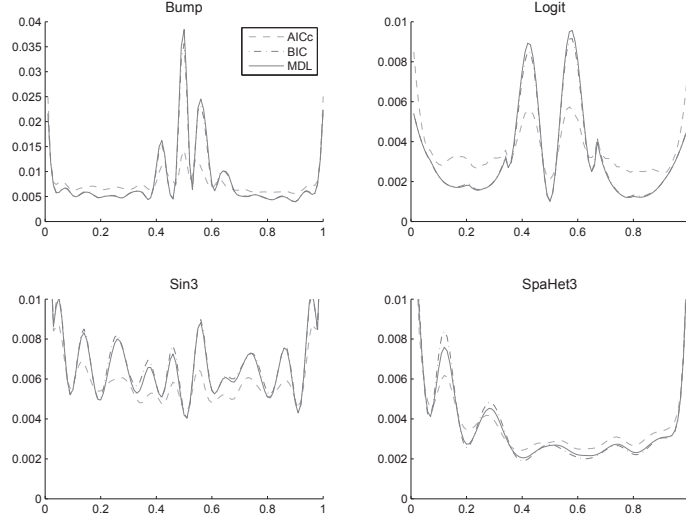


Fig. 1.4. $MSE(x)$ for each criterion as $n = 200$ (Degrees as in Table 1).

1.7. APPENDIX

The estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\delta}}, \hat{\sigma}^2)$ is a sufficient statistic for $\boldsymbol{\theta} = (\boldsymbol{\delta}, \sigma^2)$ under the model (1.17). By sufficiency the density (1.17) can be written as

$$f(\mathbf{y}; \boldsymbol{\theta}) = f(\mathbf{y}|\hat{\boldsymbol{\theta}})g(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}), \quad (1.36)$$

where the conditional density $f(\mathbf{y}|\hat{\boldsymbol{\theta}})$ does not depend on the unknown parameter vector $\boldsymbol{\theta}$. The estimators $\hat{\boldsymbol{\delta}}$ and $\hat{\sigma}^2$ are not independent like in the ordinary linear regression, but we use the approximation

$$g(\hat{\boldsymbol{\delta}}, \hat{\sigma}^2; \boldsymbol{\delta}, \sigma^2) \approx g_1(\hat{\boldsymbol{\delta}}; \boldsymbol{\delta}, \sigma^2)g_2(\hat{\sigma}^2; \sigma^2), \quad (1.37)$$

where $g_1(\hat{\boldsymbol{\delta}}; \boldsymbol{\delta}, \sigma^2)$ is the density function for the normal distribution (1.13).

The quadratic form $n\hat{\sigma}^2/\sigma^2 = \mathbf{y}'(\mathbf{I} - \mathbf{H})^2\mathbf{y}/\sigma^2$ does not follow a χ^2 -distribution, since the matrix \mathbf{H} in (1.16) is not idempotent (See e.g. Ref. 16). Let χ_ν^2 denotes a gamma variable with parameters $\nu/2$ and 2. The simple Patnaik's two-moment approximation¹⁷ consists of replacing the distribution of $Q = n\hat{\sigma}^2/\sigma^2$ by that of $c\chi_\nu^2$, where c and ν are chosen so that

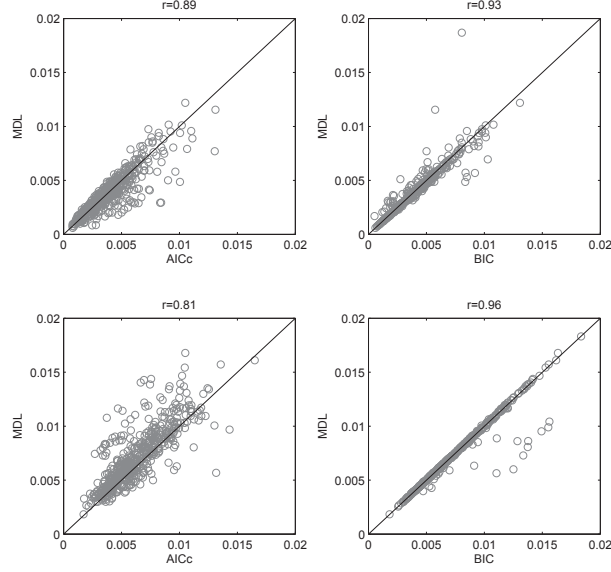


Fig. 1.5. Scatterplots of the ASE values (1.35) for SpaHet3 (Panels 1 and 2) and Sin3 (Panels 3 and 4), $n = 200$.

Q and $c\chi_\nu^2$ have the same first two moments, that is,

$$E(Q) = E(c\chi_\nu^2) \text{ and } \text{Var}(Q) = \text{Var}(c\chi_\nu^2).$$

Here ν can be fractional, and consequently χ_ν^2 is not a proper χ^2 -distribution.

However, instead of Patnaik's approximation we replace the distribution of Q by that of χ_d^2 , which has a gamma distribution with parameters $d/2$ and 2 with $d = \text{tr } \mathbf{H}$. This approximation gives results similar to Patnaik's approximation, but the derivation of the MDL criterion can be simplified. Now the approximate density g_2 of $\hat{\sigma}^2$ can be written as

$$g_2(\hat{\sigma}^2; \sigma^2) = \frac{n^{\frac{n-d}{2}}}{\Gamma(\frac{n-d}{2}) 2^{\frac{n-d}{2}}} (\hat{\sigma}^2 / \sigma^2)^{\frac{n-d}{2}} (\hat{\sigma}^2)^{-1} e^{-\frac{n\hat{\sigma}^2}{2\sigma^2}}. \quad (1.38)$$

Note that the function $\max_{\delta} g_1(\hat{\delta}; \delta, \sigma^2) \equiv \tilde{g}_1(\sigma^2)$ depends on the parameter σ^2 only. We use the approximation $\tilde{g}_1(\sigma^2) g_2(\hat{\sigma}^2; \hat{\sigma}^2) \equiv \tilde{g}(\hat{\sigma}^2)$ to

the function $g(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}})$. The function $\tilde{g}(\hat{\sigma}^2)$ can be written as

$$\tilde{g}(\hat{\sigma}^2) = A_{d,k}(\hat{\sigma}^2)^{-\frac{k}{2}-1}$$

where

$$A_{d,k} = \frac{|\mathbf{M}'\mathbf{M}|^{1/2}}{(\pi n)^{k/2}|\mathbf{B}|} \frac{\left(\frac{n}{2}\right)^{\frac{k-d}{2}}}{\Gamma\left(\frac{n-d}{2}\right)} \left(\frac{n}{2e}\right)^{\frac{n}{2}}$$

and $k = p + m$.

Utilizing the factorization (1.36) and the above approximations we get the normalizing constant in (1.29) as follows

$$\begin{aligned} C(s, R) &= \int_{\mathcal{T}(s,R)} \left[\int_{\mathcal{Y}(\hat{\boldsymbol{\theta}})} f(\mathbf{y}|\hat{\boldsymbol{\theta}}) d\mathbf{y} \right] \tilde{g}(\hat{\sigma}^2) d\hat{\boldsymbol{\theta}} \\ &= A_{d,k} \int_s^\infty (\hat{\sigma}^2)^{-\frac{k}{2}-1} d\hat{\sigma}^2 \int_{\mathcal{D}(R)} d\hat{\boldsymbol{\delta}} \\ &= A_{d,k} V_k \frac{2}{k} \left(\frac{R}{s}\right)^{k/2}, \end{aligned} \quad (1.39)$$

where $\mathcal{T}(s, R) = \{\hat{\boldsymbol{\theta}} : \hat{\sigma}^2 \geq s, \hat{\boldsymbol{\delta}}'(\mathbf{B}')^{-1}\mathbf{M}'\mathbf{M}(\mathbf{B})^{-1}\hat{\boldsymbol{\delta}} \leq a_{n,d}R\}$ is the constrained estimation space. Integrating the inner integral in the first line of (1.39) over $\mathcal{Y}(\hat{\boldsymbol{\theta}}) = \{\mathbf{y} : \hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{y})\}$ for a fixed value of $\hat{\boldsymbol{\theta}}$ gives unity. In the last line of (1.39)

$$V_k R^{k/2} = \frac{(\pi n)^{k/2} R^{k/2} |\mathbf{B}|}{\frac{k}{2} \Gamma\left(\frac{d}{2}\right) |\mathbf{M}'\mathbf{M}|^{1/2} \left(\frac{n}{2}\right)^{\frac{k-d}{2}}}$$

is the volume of an ellipsoid $\mathcal{D}(R) = \{\hat{\boldsymbol{\delta}} : \hat{\boldsymbol{\delta}}'(\mathbf{B}')^{-1}\mathbf{M}'\mathbf{M}(\mathbf{B})^{-1}\hat{\boldsymbol{\delta}} \leq a_{n,d}R\}$ ¹⁸, where $a_{n,d} = n^{k/2} \Gamma(k/2) / [(\frac{n}{2})^{\frac{k-d}{2}} \Gamma(d/2)]$. Note that $\mathcal{D}(R) = \{\tilde{\boldsymbol{\delta}} : \tilde{\boldsymbol{\delta}}' \mathbf{M}' \mathbf{M} \tilde{\boldsymbol{\delta}} \leq a_{n,d} R\}$, since $\hat{\boldsymbol{\delta}} = \tilde{\boldsymbol{\delta}} \mathbf{B}$ by (1.12). By using Rissanen's renormalization technique³ we get rid of the two parameters R and s and obtain the MDL criterion (1.30).

References

1. Rissanen, J. (1996). Fisher Information and Stochastic Complexity. *IEEE Transactions on Information Theory*, IT-42, No. 1, 40–47.
2. Rissanen, J. (2000). MDL Denoising. *IEEE Trans. on Information Theory*, IT-46, No. 1, 2537–2543.
3. Rissanen, J. (2007). *Information and Complexity and in Statistical Modeling*. New York, Springer.

4. Brumback, B. A., Ruppert, D. and Wand, M. B. (1999). Comment on Shively, Kohn and Wood. *Journal of the American Statistical Association*, 94, 794–797.
5. Ruppert, D., Wand, M. P., Carroll, R. J. (2003). *Semiparametric regression*, Wiley.
6. Demidenko, E. (2004). *Mixed models*, Wiley.
7. Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance Components*. New York, Wiley.
8. Pinheiro, J. C. and Bates, B. M. (2000). *Mixed-Effects Models in S and S-PLUS* New York, Springer.
9. Henderson, C. R. (1963). Genetic index and expected genetic advance. In *Statistical genetics and plant breeding* (W. D. Hanson and H. F. Robinsin, eds.), 141–163. National Academy of Research Council Publication No. 982, Washington, D. C.
10. Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science*, 6, 15–51.
11. Barron, A. R., Rissanen, J. and Yu, B. (1998). The MDL principle in modeling and coding. *Special Issue of Information Theory to Commemorate 50 Years of Information Theory*, 44, 2743–2760.
12. Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11, 735–754.
13. Green, D. J. and Silverman B. W. (1996). *Nonparametric regression and generalized linear models*, Chapman & Hall, London.
14. Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11, 89–121.
15. Hurvich, C. M., Simonoff, J. S., Tsai, C. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society*, B 60, 271–293.
16. Searle, S. R. (1971). *Linear Models*. New York, Wiley.
17. Patnaik, P. B. (1949). The non-central χ^2 and F -distributions and their applications. *Biometrika*, 36, 202–232.
18. Cramer, H. (1946) *Mathematical Methods of Statistics* Princeton, Princeton University Press.

- [7] Liski, A., Tabus, I., Sund, R. and Häkkinen, U. (2012)
**Variable selection by sNML criterion in
logistic regression with an application to a
risk-adjustment model for hip fracture
mortality**
Journal of Data Science 10(2), pp.321–343.

Variable Selection by sNML Criterion in Logistic Regression with an Application to a Risk-Adjustment Model for Hip Fracture Mortality

Antti Liski^{1*}, Ioan Tăbuș¹, Reijo Sund² and Unto Häkkinen²

¹*Tampere University of Technology and*

²*National Institute for Health and Welfare*

Abstract: When comparing the performance of health care providers, it is important that the effect of such factors that have an unwanted effect on the performance indicator (eg. mortality) is ruled out. In register based studies randomization is out of question. We develop a risk adjustment model for hip fracture mortality in Finland by using logistic regression. The model is used to study the impact of the length of the register follow-up period on adjusting the performance indicator for a set of comorbidities. The comorbidities are congestive heart failure, cancer and diabetes. We also introduce an implementation of the minimum description length (MDL) principle for model selection in logistic regression. This is done by using the normalized maximum likelihood (NML) technique. The computational burden becomes too heavy to apply the usual NML criterion and therefore a technique based on the idea of sequentially normalized maximum likelihood (sNML) is introduced. The sNML criterion can be evaluated efficiently also for large models with large amounts of data. The results given by sNML are then compared to the corresponding results given by the traditional AIC and BIC model selection criteria. All three comorbidities have clearly an effect on hip fracture mortality. The results indicate that for congestive heart failure all available medical history should be used, while for cancer it is enough to use only records from half a year before the fracture. For diabetes the choice of time period is not as clear, but using records from three years before the fracture seems to be a reasonable choice.

Key words: Code length, hip fracture, logistic regression, maximum likelihood.

1. Introduction

*Corresponding author.

Profiling medical care providers on the basis of quality of care and utilization of resources has become a widely used analysis in health care policy and research. A major initiative to evaluate hospital performance in the United States was launched by the Health Care Financing Administration (HCFA) in 1987 with the annual release of hospital-specific data comprising observed and expected mortality rates for Medicare patients. Hospitals observed to have higher-than-expected mortality rates were flagged as institutions with potential quality problems. HCFA derived mortality rates by estimating a patient-level model of mortality for disease-based cohorts using administrative data (Normand, Glickman and Gatsonis, 1997).

Risk-adjustment is desirable when comparing hospitals or hospital districts with respect to a performance indicator such as mortality. Adjustment is intended to account for possible differences in patient case mix (Iezzoni, 1994; Landon, Iezzoni, Ash, Shwartz, Daley, Hughes and Mackiernan, 1996; Salem-Schatz, Moore, Rucker and Pearson, 1994). The methodologic aspects of risk-adjustment have been extensively discussed in the literature on observational studies (see Rosenbaum, 2002 and references therein).

While using administrative register-based data, the comorbidities to be adjusted are typically identified from the data using the disease grouping rules defined in Charlson or Elixhauser indices (Quan, Sundararajan, Halfon, Fong, Burnand, Luthi, Saunders, Beck, Feasby and Ghali, 2005). A salient issue in adjusting performance indicators for patients' comorbidities using administrative data is to decide the length of comorbidity lookup period, i.e. to decide how far we have to go back in patient's history (recorded in the registers) in order to effectively identify comorbidities to be adjusted (Preen, Holman, Spilsbury, Semmens and Brameld, 2006). This is an important question, because all conditions might not affect the patient anymore after a certain amount of time has passed. Therefore looking back too far for a certain condition, might even make the adjustment worse. Another reason is the fact that we might have only a few years historical data available or that it is very costly to collect additional historical data. It is not desirable to collect expensive extra data if we get the same results with less information.

Often the evaluation of a risk-adjustment model for a binary response is done using the c-statistic (Iezzoni, 2003). In this approach, the probabilities estimated (typically) with logistic regression are used to predict a patient's status and the c-statistic measuring the concordance of predictions with the true events is calculated. However, accurate or inaccurate classification by c-statistic does not address the goodness of fit or the complexity of a (risk-adjustment) model (Hosmer and Lemeshow, 2000, Chapter 5). Even if the model is the correct one and thus fits very well, its classification performance may be poor. On the other hand,

the correct model may have bad fit (distances between certain observed and expected values are large) but the model still yields good classification. Clearly the aim in deciding the length of lookup period is not to find the best prediction for a single performance indicator in one data set, but to find good risk-adjustment models for further analysis. In this sense, the real model selection criteria provided should be used instead of c-statistics.

There are several traditional model selection criteria available, such as the Akaike (AIC) and the Bayesian (BIC) information criteria. Rissanen (1996) has proposed the so called minimum description length (MDL) principle which can be implemented through the normalized maximum likelihood (NML) framework. The NML distributions offer a philosophically superior approach for the model selection problem. Unfortunately, the implementation of the MDL principle for the model selection problem in logistic regression using the standard normalized maximum likelihood (NML) technique is computationally infeasible with large data sets.

This paper has two purposes. First, it develops a risk adjustment model for a binary response using logistic regression and examines the impact of the length of the register follow-up period on adjusting the performance indicator for a set of comorbidities. The second purpose of this paper is to introduce a new MDL-based model selection criterion following the idea of sequentially normalized maximum likelihood (sNML) that was recently proposed by Rissanen and Roos (2007). We show that the sNML criterion can be evaluated efficiently and it is applicable also to large models with large amounts of data by applying this criterion in the case of a risk-adjustment model for hip fracture mortality in Finland. In this case study, the focus is on the determination of the optimal length of the register follow-up periods for comorbidities. We also compare the results given by the sNML criterion with the corresponding results given by the traditional AIC and BIC model selection criteria.

2. Setting

2.1 Hip Fracture

Hip fracture is a common and important cause of mortality in the elderly population (Keene, Parker and Pryor, 1993). In Finland, the number of hip fractures was about 7000 per year between the years 1998-2002 (Sund, 2007). Not only patients suffer from hip fractures, but they also cause remarkable costs to society (Hannan, Magaziner, Wang, Eastwood, Silberzweig, Gilbert, Morrison, McLaughlin, Orosz and Siu, 2001).

The main objective in the treatment of hip fracture is to help the patient regain his/her pre-fracture health status and level of functional ability. Because a

successful treatment should make it possible that patients are able to continue life in the same fashion as before the fracture, death is obviously a very unsuccessful outcome. Although hip fracture itself doesn't usually cause death, it is often such a shock to the whole body that especially for elderly people in lowered physical state it may mean the "beginning of the end" (Heithoff and Lohr, 1990). If the hip fracture triggers the dying process, we may assume that short-term mortality is in fact an indicator that the patient's health status before the hip fracture was already substantially lowered.

Quite often the mortality indicators for hip fracture are selected to measure death within three months or one year after the fracture. Mortality is a well defined and easily observable indicator in the sense that there is typically no argument if a patient is dead or not. The 90 days mortality reflects the risk connected to the hip fracture treatment and one year mortality reflects more the overall condition of a patient than risk of death directly caused by the shock effect of the hip fracture event.

2.2 Adjusting Mortality with Comorbidities

In order to compare mortality indicators between different areas or in time, the differences or changes in the patient population must be risk-adjusted (Iezzoni, 2003). In other words, we wish to find factors that explain the mortality following hip fracture, measured as a binary variable, in order to obtain a set of covariates which profile a patient's medical condition at the time of the hip fracture. Our interest is in comorbidities that a patient has had before the hip fracture and which may have effect on the outcome of the treatment. The special focus in our study is to examine how far we have to follow the patients medical history, and various lengths of the follow-up period (180 days, 1 year, 3-, 5- and 10-years) are modeled in order to find the shortest period to effectively adjust for each comorbidity. For pragmatic reasons, only three comorbidities are used in this study: congestive heart failure, cancer and diabetes. Each time period and comorbidity is analyzed separately. The analysis for other comorbidities could be done in a similar fashion. On top of the comorbidities, age, hip fracture type and sex are considered as factors to be adjusted in our risk-adjustment model.

2.3 Data

The National Institute for Health and Welfare maintains a register which contains all in hospital care periods taken place in Finland. From this register all 50 year or older first time hip fracture patients were identified from the years 1999 – 2005. We further excluded patients who were institutionalized before the fracture. This resulted in a total of $n = 28797$ patients. For these patients (back-

wards) hospitalization history was available up to 10 years before the fracture. This information was complemented with data obtained from the register maintained by the Social Insurance Institution of Finland. From this second register, information on drug reimbursements granted for the medication of the three comorbidities stated above, was obtained. The mortality was followed using the Causes of Death register of Statistics Finland. In our final data we have combined the information obtained from these three registers. It has been shown that the quality of Finnish register data on the case of hip fractures is good (Sund, Nurmi-Lüthje, Lüthje, Tanninen, Narinen and Keskimäki, 2007). The dataset is based on the data used in the PERFECT (PERFormance, Effectiveness and Cost of Treatment episodes) project in the National Institute for Health and Welfare in Finland.

Many basic characteristics can be straightforwardly extracted from the data. These include the date of hip fracture, sex, age, the type of hip fracture (subtrochanteric, trochanteric or femoral neck fracture), and the date of death. In addition, we used ten years of medical history to construct five variables for each comorbidity which scan different time periods before hip fracture. The time intervals of interest were 180 days, 1 year, 3 years, 5 years and 10 years before the fracture. There were two ways to get an indication for a comorbidity from our data. In the first we have data on a patient's all hospitalization preceding the hip fracture until a certain (historical) time point. Now if the patient has been hospitalized because of the chosen comorbidity between this time point and the hip fracture, we get indication that the patient has had that comorbidity. The second way to get indication for a comorbidity comes through information on drug reimbursements. Now we have to check if a patient has received the right for drug reimbursements for that comorbidity and that it was still valid when the hip fracture occurred. This means that if a patient has had the right for drug reimbursements when the hip fracture occurred, then the patient will have indication for that comorbidity for all time periods.

Let us take an example where we choose the 3 year time interval. This means we jump back three years in time from the hip fracture event. We now choose one patient whose hospitalization record we start following towards the hip fracture event. Assume the patient has been hospitalized because of cancer for three weeks two years before the hip fracture. Now this patient will be identified for cancer based on the information on hospital care records. It is also checked if the patient has a right for drug reimbursements for some of the three comorbidities that we are interested in at the moment of hip fracture. Say we find out that the patient has the right for drug reimbursements because of cancer but also for congestive heart failure. Therefore this patient receives indication for cancer (based on information from both registers) and congestive heart failure with a

three year lookback period.

The setting is actually quite challenging from the model selection point of view, since the number of the occurrences of a disease does not increase much when the length of inspection period increases. If we change our view for example from 180 days to one year before the fracture, the increase in the number of occurrences is typically small. Therefore it may be difficult to distinguish between models that use different time period variables. Further, if we look further back in history, more occurrences appear, but the effect of these occurrences on the dependent variable may become weaker, and we assume that this time dependence may not be same for all comorbidities.

3. Modeling Mortality with Logistic Regression

With n patients, we define $y_t = 1$ if the t th patient died within a 90 days period after the hip fracture and $y_t = 0$ otherwise (A corresponding model for the 365 days mortality is also analysed). We treat the n binary outcome variables y_1, \dots, y_n as independent. Let

$$\pi(\mathbf{x}_t; \boldsymbol{\beta}) = P(y_t = 1), \quad t = 1, \dots, n,$$

and assume that

$$\log \frac{\pi(\mathbf{x}_t; \boldsymbol{\beta})}{1 - \pi(\mathbf{x}_t; \boldsymbol{\beta})} = \boldsymbol{\beta}^T \mathbf{x}_t, \quad (1)$$

where $\mathbf{x}_t = (x_{t1}, \dots, x_{tk})^T$ is the vector of k covariate values of the t th patient.

The covariates (comorbidity, age, sex and hip fracture type) are such that there are many patients with the same values of covariates. For example, we may take women patients in the age group 50-69 who suffered a subtrochanteric fracture and had diabetes (inspection period one year before fracture). Let n_1 denote the number of such patients. Consequently, these patients have the same value of covariates, say \mathbf{x}_1 , and hence the probability $P(y_t = 1)$ is $\pi(\mathbf{x}_1; \boldsymbol{\beta})$ for all these n_1 patients. We say that \mathbf{x}_1 is the setting 1 of values of k covariates. We have only l different settings $\mathbf{x}_1, \dots, \mathbf{x}_l$ and the number of different setting l is much smaller than n . Let n_i denote the number of the patients with the setting i , and hence we have $n = n_1 + \dots + n_l$.

3.1 Bernoulli Likelihood

For notational convenience, we assume here that the observations are ordered such that the different settings $\mathbf{x}_1, \dots, \mathbf{x}_l$ come first, i.e. for each $t > l$ there exists $i \leq l$ such that $\mathbf{x}_t = \mathbf{x}_i$. Since the y_t are independent and Bernoulli distributed,

the likelihood is

$$\begin{aligned} L(\beta; \mathbf{x}_1, \dots, \mathbf{x}_n) &= \prod_{t=1}^n \pi(\mathbf{x}_t; \beta)^{y_t} [1 - \pi(\mathbf{x}_t; \beta)]^{1-y_t} \\ &= \prod_{i=1}^l \pi(\mathbf{x}_i)^{v_i} [1 - \pi(\mathbf{x}_i)]^{n_i - v_i}, \end{aligned} \quad (2)$$

where $v_i = \sum_{t: \mathbf{x}_t = \mathbf{x}_i} y_t$ is the number of deaths among the patients with the setting \mathbf{x}_i , $i = 1, \dots, l$. Therefore it is sufficient to record the number of observations n_i and the number of deaths v_i corresponding to the settings $i = 1, \dots, l$. Then v_i refers to this death count rather than to an individual binary response. We will use logistic regression (DeLong *et al.*, 1997) to assess from register data how much of medical history before fracture is needed in order to get sufficient indication of comorbidity effects.

3.2 Model Selection in Logistic Regression

Let Γ be the set of all $1 \times k$ vectors of the form $\gamma = (\gamma_1, \dots, \gamma_k)$, where $\gamma_j = 0$ or 1 for $j = 1, \dots, k$. There are 2^k such vectors in Γ . A variable selection procedure is then equivalent to first selecting $\gamma \in \Gamma$. If $\gamma_j = 1$, the variable x_j , $1 \leq j \leq k$ is selected and the corresponding β_j is estimated, otherwise $\gamma_j = 0$ and $\beta_j = 0$, i.e. x_i is not selected. Let $\beta_\gamma = \text{diag}[\gamma]\beta$, where $\text{diag}[\gamma]$ is the $k \times k$ diagonal matrix with diagonal elements γ and $\beta = (\beta_1, \dots, \beta_k)^T$ is the k -dimensional parameter vector. In our application we will consider a certain subset of models from Γ (See Section 4.1) and compare them using the model selection criteria NML, AIC and BIC.

It follows from assumption (1) and the likelihood (2) that the log likelihood function of β_γ equals

$$l(\beta_\gamma) = \sum_{i=1}^l v_i \beta_\gamma^T \mathbf{x}_i - \sum_{i=1}^l n_i \log[1 + \exp(\beta_\gamma^T \mathbf{x}_i)]. \quad (3)$$

The likelihood equations result from setting $\partial l(\beta_\gamma)/\partial \beta_\gamma = 0$, and they may be written in the form

$$\mathbf{X}^T \mathbf{v} = \mathbf{X}^T \hat{\boldsymbol{\mu}},$$

where $\mathbf{v} = (v_1, \dots, v_l)^T$, $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_l)^T$ and $\hat{\mu}_i = n_i \pi(\mathbf{x}_i; \hat{\beta}_\gamma)$, $i = 1, \dots, l$. The equations are nonlinear and require iterative solution. The likelihood equations equate the sufficient statistics to the estimate of their expected values. This is a fundamental result for generalized linear models with canonical link (see eg. McCulloch and Searle 2001, Chapter 5).

4. The MDL Principle and the NML Criterion for Logistic Regression

4.1 Normalized Maximum Likelihood

Rissanen (1996) proposed his normalized maximum likelihood (NML) distribution as a theoretical basis for statistical modeling. The NML distribution for (2) may now be written as

$$\hat{P}(\mathbf{v}|\gamma) = L[\hat{\beta}_\gamma(\mathbf{v}) | \mathbf{v}; \mathbf{X}] / C(\gamma), \quad (4)$$

where $L[\hat{\beta}_\gamma(\mathbf{v}) | \mathbf{v}; \mathbf{X}]$ is the maximum of the likelihood function and

$$C(\gamma) = \sum_{\mathbf{v} \in \Omega} L[\hat{\beta}_\gamma(\mathbf{v}) | \mathbf{v}; \mathbf{X}] \quad (5)$$

is the normalizing constant. In (5) Ω denotes the sample space and the sum runs over all different count vectors (v_1, \dots, v_l) such that $0 \leq v_1 + \dots + v_l \leq n$ and $v_i \geq 0$, $i = 1, \dots, l$. The notation $\hat{\beta}_\gamma(\mathbf{v})$ emphasizes the obvious fact that the ML estimate $\hat{\beta}_\gamma$ is a function of \mathbf{v} .

There is a correspondence between so called prefix codes and probability distributions (Rissanen, 2007, Chapter 2). Let $P(\mathbf{v} | \beta_\gamma)$ be the probability of \mathbf{v} . Then there exists a prefix code for \mathbf{v} with ideal code length $\log[1/P(\mathbf{v} | \beta_\gamma)] = -\log P(\mathbf{v} | \beta_\gamma)$. So, every distribution defines a prefix code. After observing \mathbf{v} , the shortest code length is $\log(1/P[\mathbf{v} | \hat{\beta}_\gamma(\mathbf{v})])$. Clearly the maximum likelihood $P[\mathbf{v} | \hat{\beta}_\gamma(\mathbf{v})]$ is not a probability distribution of \mathbf{v} , and therefore it does not define a prefix code for \mathbf{v} . However, the NML distribution (4) defines a prefix code which has important optimality properties (see eg. Barron, Rissanen and Yu, 1998).

4.2 The Minimum Description Length Principle

Rissanen (1996) considers the NML distribution in the context of coding and modeling theory and takes

$$-\log \hat{P}(\mathbf{v}|\gamma) = -l[\hat{\beta}_\gamma(\mathbf{v}) | \mathbf{v}; \mathbf{X}] + \log C(\gamma) \quad (6)$$

as the “shortest code length” for the data \mathbf{v} that can be obtained with the model γ and calls it the stochastic complexity of \mathbf{v} , given γ . The first term in (6) is the minimized negative log likelihood, and the second term is called parametric complexity. In essence, $-\log \hat{P}(\mathbf{v}|\gamma)$ is the minimum of the penalized log likelihood function. The minimized negative log likelihood measures goodness of fit to the data, while $\log C(\gamma)$ penalizes the complexity of the model γ . From the coding

theoretic point of view, $-\log \hat{P}(\mathbf{v}|\gamma)$ is the length of the prefix code defined by the NML distribution.

Here we consider the class of logistic regression models defined by the 2^k subsets of covariates Γ and the logistic probabilities. The aim of model selection is to pick the optimal model γ from the set Γ . For given data \mathbf{v} , the NML function (4) attains its maximum and the “code length” (6) its minimum at the same value of γ . According to *the MDL* (Minimum Description Length) principle (Rissanen, 2007, Chapter 8) we select the model $\hat{\gamma}$ from Γ that minimizes the stochastic complexity (6). Since $\hat{\gamma}$ maximizes (4), we may call it the NML estimate of γ within the model class Γ .

The code length interpretation of (6) provides an illustrative yardstick to compare models. The data can be considered as a sequence of zeros and ones 0010100...0010, where 1 refers to “death”. The upper limit of the code length is the length $n = 28797$ of the whole sequence. If a model will capture the regular features of data well, then the prefix code based on the NML distribution (4) can compress the data sequence. Our optimal logistic regression risk adjustment model compresses the data sequence into a sequence whose length is about half of the upper limit 28797. No actual coding is needed, of course, but the stochastic complexity of a model is computed.

Unfortunately, the computational burden becomes too heavy to determine the value of $C(\gamma)$ for logistic regression models with moderate number of covariates when n is large. Let k_γ denote the number of covariates in the model γ and l_γ the number of different settings of covariate values in the data under the model γ . Then the sum in (5) runs over all different count vectors $(v_1, \dots, v_{l_\gamma})$ such that $0 \leq v_1 + \dots + v_{l_\gamma} = v_\gamma \leq n$ and $0 \leq v_i \leq n_i$, $i = 1, \dots, l_\gamma$, where $n = 28797$. Let γ be a model with two covariates ($k_\gamma = 2$), say. When the covariates are dichotomous, there are 2^2 possible covariate settings. Suppose that in the data occur only the settings $(0, 0)$, $(1, 0)$ and $(0, 1)$, and hence $l_\gamma = 3$. Then v_γ takes the values $0, 1, \dots, n$ and for each v_γ the count vectors are obtained by determining all different partitionings of v_γ into (v_1, v_2, v_3) such that $v_1 + v_2 + v_3 = v_\gamma$, $0 \leq v_i \leq n_i$, $i = 1, 2, 3$ and $n_1 + n_2 + n_3 = n$. The ML estimate has to be computed for each count vector. It is obvious that the computation of the code length for just one model is excessive not to mention the situation where we wish to compare several models.

Tabus and Rissanen (2006) presented an algorithm for the computation of the stochastic complexity (6) for logistic regression. If the number of covariates is $k = 3$, say, their algorithm is practical only in cases with a maximum of a few hundred observations. The sequentially normalized ML technique will decrease computational burden dramatically, and consequently it makes the MDL model selection practical also for models with large k and n .

4.3 Sequential NML

The sequentially normalized maximum likelihood was introduced by Roos and Rissanen (2007). This approach has the advantage that the normalizing constant is much easier to compute than in the case of the standard NML. Now we only need to normalize over the last observation, which simplifies computations substantially. On the other hand, if we don't have a strict order for the data, we have to choose one, and this ordering has naturally an effect on the results.

Roos and Rissanen (2008, equation 4) presented the sequentially normalized maximum likelihood (sNML) function. Let $X^n = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ denote the regressor matrix and $\mathbf{y}^n = (y_1, \dots, y_n)$ a sequence of the binary outcome variables. Note that here \mathbf{x}_i denotes the regressor vector of the i th patient and X^n may contain identical regressor vectors unlike \mathbf{X} in the model described in (3). In the logistic regression case, the sNML function may be written as

$$\hat{P}(\mathbf{y}^n | X^n) = \hat{P}(\mathbf{y}^m | X^m) \prod_{t=m+1}^n \hat{P}(y_t | \mathbf{y}^{t-1}, X^t), \quad (7)$$

where $\hat{P}(\mathbf{y}^n | X^n)$ is the estimated probability to observe the string \mathbf{y}^n having observed X^n .

The last term from (7) is the NML function for y_t

$$\hat{P}(y_t | \mathbf{y}^{t-1}, X^t) = \frac{P(y_t | \mathbf{y}^{t-1}, X^t, \hat{\boldsymbol{\beta}}(\mathbf{y}^t))}{K(\mathbf{y}^{t-1})}, \quad (8)$$

where

$$K(\mathbf{y}^{t-1}) = P(y_t = 0 | \mathbf{y}^{t-1}, X^t, \hat{\boldsymbol{\beta}}_0) + P(y_t = 1 | \mathbf{y}^{t-1}, X^t, \hat{\boldsymbol{\beta}}_1) \quad (9)$$

is the normalizing constant.

Here $\hat{\boldsymbol{\beta}}_i$, denotes the ML estimates of $\boldsymbol{\beta}$ from the binary outcome vector (\mathbf{y}^{t-1}, i) , $i = 0, 1$ respectively. As can be seen from (8) we only normalize over the last observation which simplifies the computation of the normalizing constant compared to the standard NML.

Because the observations are independent, we have

$$P(y_t = i | \mathbf{y}^{t-1}, X^t, \hat{\boldsymbol{\beta}}_i) = \frac{(e^{\hat{\boldsymbol{\beta}}_i^T \mathbf{x}_t})^i}{1 + e^{\hat{\boldsymbol{\beta}}_i^T \mathbf{x}_t}}, \quad i = 0, 1,$$

and (8) becomes

$$\hat{P}(y_t = i | \mathbf{y}^{t-1}, X^t) = \left(\frac{(e^{\hat{\boldsymbol{\beta}}_i^T \mathbf{x}_t})^i}{1 + e^{\hat{\boldsymbol{\beta}}_i^T \mathbf{x}_t}} \right) / \left(\frac{1}{1 + e^{\hat{\boldsymbol{\beta}}_0^T \mathbf{x}_t}} + \frac{e^{\hat{\boldsymbol{\beta}}_1^T \mathbf{x}_t}}{1 + e^{\hat{\boldsymbol{\beta}}_1^T \mathbf{x}_t}} \right), \quad i = 0, 1.$$

The negative logarithm of the sNML function (7) is

$$\begin{aligned}
 -\log \hat{P}(\mathbf{y}^n | X^n) &= -\log \hat{P}(\mathbf{y}^m | X^m) - \sum_{t=m+1}^n \log \hat{P}(y_t | \mathbf{y}^{t-1}, X^t) \\
 &= -\log \hat{P}(\mathbf{y}^m | X^m) - \sum_{t=m+1}^n \log P(y_t | \mathbf{y}^{t-1}, X^t, \hat{\beta}(\mathbf{y}^t)) \\
 &\quad + \sum_{t=m+1}^n \log K(\mathbf{y}^{t-1}). \tag{10}
 \end{aligned}$$

The computational burden of $\sum_{t=m+1}^n \log K(\mathbf{y}^{t-1})$ in (10) is trivial contrary to the computation of $\log[C(\gamma)]$ in (4). Note that $-\log \hat{P}(\mathbf{y}^n | X^n)$ can be interpreted as the code length for data when a given model is used, as explained in Subsection 4.2.

4.4 Individual Code Lengths

Taking the negative base two logarithm of (7), yields

$$-\log_2 \hat{P}(\mathbf{y}^n | X^n) = -\log_2 \hat{P}(\mathbf{y}^m | X^m) - \sum_{t=m+1}^n \log_2 [\hat{P}(y_t | \mathbf{y}^{t-1}, X^t)],$$

where the last term is just a sum of the code lengths of individual observations from $m+1$ to n . Thus we are able to consider the contribution of individual observations to the total code length. Let \mathcal{S} denote a subsequence s_1, s_2, \dots, s_v of the sequence $m+1, m+2, \dots, n$ of indices. Thus $m+1 \leq s_1 < s_2 < \dots < s_v \leq n$, where $v \leq n-m$ is the number of indices in \mathcal{S} . Next take a single index $s \in \mathcal{S}$ and let X^s denote the sequence $(\mathbf{x}_1, \dots, \mathbf{x}_m, x_{m+1}, \dots, x_{s-1}, x_s)$. The sequence X^s has s elements, $m+1 \leq s \leq n$. In a similar fashion \mathbf{y}^s denotes the corresponding sequence of binary outcomes $(y_1, \dots, y_m, y_{m+1}, \dots, y_s)$.

We may now compare how changing explanatory variables affects the code length. Let X_1 and X_2 be two different sets of explanatory variables. The change in code length y_s (or description for y_s) is obtained by computing

$$\log_2 [\hat{P}(y_s | \mathbf{y}^{s-1}, X_2^s)] - \log_2 [\hat{P}(y_s | \mathbf{y}^{s-1}, X_1^s)].$$

By summing up these individual differences over \mathcal{S} we obtain

$$d_S(X_1, X_2) = \sum_{s \in \mathcal{S}} \{\log_2 [\hat{P}(y_s | \mathbf{y}^{s-1}, X_2^s)] - \log_2 [\hat{P}(y_s | \mathbf{y}^{s-1}, X_1^s)]\}, \tag{11}$$

which tells us how much the observations belonging to \mathcal{S} affect the total code length as we switch our set of explanatory variables from X_2 to X_1 . Let X_1^C be the comorbidity variable “cancer” using one year of a patient’s medical history, and $X_{1/2}^C$ the corresponding variable using a half year medical history. Then $d_S(X_1^C, X_{1/2}^C)$ gives the change of the code length when the patients belong to the set \mathcal{S} and one year of the medical history is used instead of half a year when cancer is the comorbidity variable. Here we may understand as well that there are also other variables in the model but only the variable X_1^C is changed to $X_{1/2}^C$.

We note that (11) is generally not the same as

$$\log_2 \hat{P}(\mathbf{y}^n | X_2^n) - \log_2 \hat{P}(\mathbf{y}^n | X_1^n). \quad (12)$$

In the case where \mathcal{S} is the full sequence of $n - m$ indices $m + 1, m + 2, \dots, n$, we have equality between (11) and (12).

4.5 Nonconstant Covariate Effects

If we assume that the effects of covariates may change over time, the calculation for the code length of each observation should be done by using an appropriately selected subdata from the near past. One choice is to slide a window over the data. The sNML approach is suitable for this purpose, although not without problems. Let us consider a window of w observations. Now to encode the whole data, we need to calculate first the ‘regular’ NML code length for the first w observations (term $\hat{P}(\mathbf{y}^m | X^m)$ in (7)). If now w is large, we face the same problems as before in the calculation of the normalizing constant (5). In the case study, we have circumvented this problem by just focusing on the comparison of the code length calculated for the patients with the indices [501, 28797]. This way we are using the information from the first 500 observations in encoding, but we do not include the cost of coding of the first 500 observations in the total code length.

5. Statistical Analysis

We analyse Finnish register data on hip fracture patients from the years 1999-2005. The data was described in Section 2.3. We have two binary outcome variables, the 90 days mortality and the 365 days mortality, with possible outcomes 1 (died) and 0 (alive). These mortality variables indicate if the patient has died within the 90 days (or the 365 days, respectively) period after the hip fracture.

The basic set of explanatory variables consists of five constructed dichotomous variables. From the national registries we have information on the hip fracture

type categorized in three classes, trochanteric, subtrochanteric and femoral neck fractures. Patients are classified into three age groups (50-69, 70-89, 90-). We use five dummy variables: two dummies for the hip fracture type, two dummies for age and one dummy for sex.

The outcome variables 90 days mortality and one year mortality will be modeled separately. The comorbidities of interest, congestive heart failure, cancer and diabetes are measured in five time intervals. Therefore we have fifteen comorbidity variables. The five basic explanatory variables are included in all models. In addition to them, one comorbidity variable is selected from the set of 15 comorbidity variables, giving 15 alternative models with six explanatory variables and a constant in each model.

We will do the analysis under two different assumptions: (1) the covariate effects change over time and (2) the covariate effects stay constant. Under the second assumption, we utilize the full medical history at each point in the computation of the sNML criterion. Under the first assumption, a sliding window technique is used.

We compute sNML with $m = 25$ (see (7)). The value $m = 25$ was chosen to make sure we have enough dead and alive patients in the initial calculation of sNML (done with the regular NML) (see Albert and Anderson, 1984). We cannot estimate a model if we only have for example dead patients in our data.

When using the sliding window, we use only a limited number of past observations to calculate the code length for an observation y_t . Now (7) becomes

$$\hat{P}(\mathbf{y}^n | X^n) = \hat{P}(\mathbf{y}^m | X^m) \prod_{t=m+1}^n \hat{P}(y_t | \mathbf{y}^{t-w-1, \dots, t-1}, X^{t-w-1, \dots, t}), \quad (13)$$

where w is the window length and $\mathbf{y}^{t-w-1, \dots, t-1} = (y_{t-w-1}, \dots, y_{t-1})$ and $X^{t-w-1, \dots, t-1}$ is the corresponding regressor matrix. In our calculations with the sliding window we take $m = 500$ and drop the term $\hat{P}(\mathbf{y}^m | X^m)$ from our code lengths because it is not possible to calculate the regular NML with 500 observations. In our setting $\hat{P}(\mathbf{y}^m | X^m)$ with $m = 500$ is always constant (or very close to constant) between the different models so omitting it doesn't really make a difference to our comparisons as they are done between different time periods of a comorbidity.

As we increase the time backwards from the hip fracture event, we get more occurrences for each comorbidity. This means that in the data some 0's of explanatory variables turn into 1's, but otherwise the data stays exactly the same when increasing the time period.

Let \mathcal{A} be the set of indices of the observations that change as we switch from the comorbidity variable X_{t_1} to X_{t_2} . The subindices t_1 and t_2 indicate the length of the time periods that we look backwards from the hip fracture

event. Now the length of time period 1 is less than time period 2. By (11), we compute $d_{\mathcal{A}}(X_{t_1}, X_{t_2})$ to obtain the change in total code length due to changing observations.

5.1 Results

We observe that in Table 1 all model selection criteria give results consistent with each other. The AIC and BIC values were calculated from the whole data (not sequentially) to show that in this case the sequential and non-sequential approaches give similar results. The formulas for AIC and BIC are

$$\text{AIC} = -2 \log P(\mathbf{y}^n | X^n; \hat{\boldsymbol{\beta}}) + 2k$$

and

$$\text{BIC} = -2 \log P(\mathbf{y}^n | X^n; \hat{\boldsymbol{\beta}}) + k \log n,$$

where k is the number of estimated parameters in the model (see eg. Burnham & Anderson, 2002, Chapter 6). These criteria can be easily calculated in the case of logistic regression model.

C-statistic values (calculated from the whole data) are reported because they are often used in this kind of analysis. Also notice that the comorbidities seem to behave quite different from each other. Congestive heart failure works best as an explanatory variable if we use all of the data available to us. Cancer is a good explanatory variable for mortality with just information from 180 days preceding the fracture. In the case of diabetes, it is difficult to distinguish between time periods. There is very little variation in the values of the model selection criteria and the time periods from three to ten years give virtually the same values. However, all the model selection criteria except c-statistic seem to make the same choice of time period also for diabetes. We notice also that our models fit better the 90 days mortality than one year mortality.

Table 2 reports the number of occurrences of congestive heart failure (CHF), cancer and diabetes in each five time periods. The increase in occurrences is not very big compared to the size of the whole data ($n=28797$). This might be the reason why the model selection criteria do not clearly prefer any model over the others. Especially this is the case with diabetes. The maximum increase of occurrences is in congestive heart failure as we extend the period from 180 days to ten years (1376 occurrences). If we don't include any of the comorbidities in the model, we get the code length (sNML) of 15113 bits for the 90 days mortality and 21339 bits for one year mortality. Even though the time periods within diabetes do not differ from each other, they all clearly improve the models compared to the models without any comorbidities.

Table 1: Code lengths (sNML), AIC, BIC and c-statistic values for 30 mortality models for each comorbidity are given. The basic variables fracture type, age and sex are included in all models and exactly one of the 15 comorbidity variables is selected for each alternative model. Models for 90 day and one year (values in brackets) mortality are given

CHF				
	sNML	AIC	BIC	c-statistic
180 days	14898 (21010)	20611 (29086)	20669 (29144)	0.6746 (0.6585)
1 year	14869 (20971)	20572 (29033)	20629 (29091)	0.6767 (0.6609)
3 years	14837 (20924)	20526 (28968)	20584 (29026)	0.6799 (0.6635)
5 years	14833 (20902)	20520 (28936)	20578 (28993)	0.6804 (0.6649)
10 years	14825 (20879)	20509 (28904)	20567 (28962)	0.6812 (0.6658)
CANCER				
	sNML	AIC	BIC	c-statistic
180 days	14980 (21011)	20719 (29079)	20777 (29137)	0.6644 (0.6599)
1 year	14979 (21013)	20721 (29085)	20779 (29143)	0.6647 (0.6560)
3 years	14992 (21026)	20738 (29103)	20796 (29160)	0.6653 (0.6550)
5 years	14991 (21062)	20765 (29152)	20823 (29210)	0.6640 (0.6538)
10 years	15005 (21067)	20757 (29160)	20815 (29218)	0.6646 (0.6540)
DIABETES				
	sNML	AIC	BIC	c-statistic
180 days	15093 (21278)	20882 (29457)	20940 (29515)	0.6549 (0.6409)
1 year	15091 (21273)	20879 (29449)	20937 (29507)	0.6553 (0.6414)
3 years	15089 (21272)	20877 (29448)	20934 (29506)	0.6557 (0.6420)
5 years	15090 (21273)	20878 (29450)	20936 (29508)	0.6560 (0.6422)
10 years	15090 (21274)	20877 (29451)	20935 (29509)	0.6561 (0.6423)

Table 2: Number of occurrences of the comorbidities within different time periods

time period	CHF	CANCER	DIABETES
180days	4654	2205	4064
1 year	4947	2470	4152
3 years	5570	2926	4305
5 years	5820	3237	4374
10 years	6030	3548	4420

In Table 3 we have reported the code lengths computed for the three comorbidities by using sliding windows of different lengths. The performance of sNML with various window lengths is close to that presented in Table 1, except for

90 days mortality with a window length of 25 observations and congestive heart failure as comorbidity. For diabetes the models with various time periods are still quite close to each other. Note, however, that Table 3 and Table 1 are not directly comparable because in the calculations for Table 3 we have omitted the code length for the first 500 observations.

Table 3: Code lengths (sNML) for 30 mortality models for each comorbidity (as in Table 1) using sliding windows of different lengths (25, 50, 100 and 500 observations). Code lengths for one year mortality are in brackets

CHF				
	25 obs	50 obs	100 obs	500 obs
180 days	18883 (23710)	17059 (22760)	15816 (21814)	14838 (20836)
1 year	18883 (23689)	17044 (22728)	15793 (21785)	14811 (20801)
3 years	18935 (23669)	17033 (22685)	15761 (21736)	14783 (20755)
5 years	18950 (23673)	17022 (22663)	15751 (21704)	14783 (20733)
10 years	18954 (23658)	17025 (22646)	15744 (21692)	14777 (20712)
CANCER				
	25 obs	50 obs	100 obs	500 obs
180 days	18676 (23571)	17098 (22640)	15901 (21785)	14913 (20818)
1 year	18760 (23625)	17111 (22644)	15907 (21782)	14915 (20817)
3 years	18850 (23680)	17147 (22671)	15914 (21801)	14927 (20835)
5 years	18916 (23717)	17187 (22729)	15932 (21832)	14945 (20869)
10 years	18967 (23733)	17192 (22741)	15934 (21854)	14945 (20877)
DIABETES				
	25 obs	50 obs	100 obs	500 obs
180 days	19169 (24047)	17335 (23006)	16028 (22078)	15030(21110)
1 year	19180 (24042)	17326 (23005)	16021 (22068)	15027(21104)
3 years	19186 (24036)	17327 (22997)	16017 (22053)	15023(21099)
5 years	19190 (24041)	17323 (22996)	16021 (22056)	15023(21097)
10 years	19193 (24042)	17329 (22999)	16027 (22062)	15022 (21095)

In Table 4 we have the change in code length within the subset of added occurrences and the whole data. With added occurrences we mean the observations that will become new occurrences of a comorbidity as we extend the time period. Let \mathcal{A} denote the set of added occurrences as in Section 5. Note that for all different pairs of time periods (in connection of a given comorbidity) we have a different set of added occurrences. For example, let $X_{t_i}^{\text{CHF}}$ be the comorbidity variable “congestive heart failure”, when the period t_i of patients medical history before hip fracture is used. If $t_1 = 1/2$ year and $t_2 = 1$ year, then by (11) $d_{\mathcal{A}}(X_{t_1}^{\text{CHF}}, X_{t_2}^{\text{CHF}})$ is the first figure (24.3097) in the first row of Table 4.

Table 4: Differences of code lengths (sNML) for changing observations and for the whole data. In the table we have $d_A(X_{t_1}, X_{t_2})$ and $d(X_{t_1}, X_{t_2})$ (see (11) and Section 5) values with different time periods for t_1 and t_2 . If for example $t_1 = 1$ year and $t_2 = 10$ years, take congestive heart failure as comorbidity and 90 days mortality as outcome, then $d_A(X_{t_1}, X_{t_2}) = 30.6144$. For the whole data $d(X_{t_1}, X_{t_2}) = 44.0480$. Values for one year mortality are in brackets

CHF			
		ch obs	all
180 days	1 year	24.3097 (34.2059)	28.2030 (38.2309)
	3 years	46.8802 (70.0761)	60.9309 (85.1012)
	5 years	48.1942 (88.2988)	64.5818 (107.7948)
	10 years	52.6179 (104.8533)	72.2511 (130.1355)
1 year	3 years	23.9742 (37.7484)	32.7278 (46.8703)
	5 years	25.5732 (56.4925)	36.3788 (69.5639)
	10 years	30.6144 (74.1631)	44.0480 (91.9046)
3 years	5 years	1.6602 (19.9655)	3.6510 (22.6937)
	10 years	6.9804 (38.7493)	11.3202 (45.0343)
5 years	10 years	5.2479 (19.2118)	7.6692 (22.3407)
CANCER			
		ch obs	all
180 days	1 year	-5.0550 (-9.0281)	0.3741 (-1.9093)
	3 years	-18.9243 (-25.5112)	-11.8064 (-14.5274)
	5 years	-35.3211 (-60.0382)	-31.4578 (-50.5906)
	10 years	-32.8788 (-65.0620)	-25.2902 (-55.9012)
1 year	3 years	-15.9025 (-20.8443)	-12.1805 (-12.6182)
	5 years	-33.0806 (-57.6566)	-31.8319 (-48.6814)
	10 years	-30.6938 (-63.7291)	-25.6643 (-53.9920)
3 years	5 years	-20.5161 (-41.9101)	-19.6514 (-36.0632)
	10 years	-18.1913 (-50.8998)	-13.4838 (-41.3738)
5 years	10 years	2.1115 (-12.5200)	6.1676 (-5.3106)
DIABETES			
		ch obs	all
180 days	1 year	2.5761 (5.4076)	2.5049 (5.3779)
	3 years	4.3044 (5.6134)	4.1464 (6.3606)
	5 years	2.6456 (3.5228)	3.3646 (4.8051)
	10 years	2.9066 (2.6675)	3.7320 (4.1115)
1 year	3 years	1.7909 (0.2711)	1.6415 (0.9828)
	5 years	0.1771 (-1.8013)	0.8597 (-0.5727)
	10 years	0.4110 (-2.6583)	1.2271 (-1.2664)
3 years	5 years	-1.6343 (-2.1389)	-0.7818 (-1.5555)
	10 years	-1.3560 (-2.9672)	-0.4144 (-2.2492)
5 years	10 years	0.3228 (-0.8389)	0.3674 (-0.6937)

If $t_1 = 5$ years and $t_2 = 10$ years, then $d_A(X_{t_1}^{\text{CHF}}, X_{t_2}^{\text{CHF}}) = 5.2479$ is the first figure in the tenth row of Table 4. It is understood here that the basic explanatory variables (hip fracture type, age and sex) and a constant are included in all models.

Table 4 shows the same tendency as the results in Table 1. We observe how much the code length changes within the subset of added occurrences and the whole data. In the case of congestive heart failure, increasing the time period shortens the code length among the added occurrences and also within the rest of the data. This means the added occurrences fit the data better with outcome value 1 than with value 0 and also improve the fit for observations coming after them.

Cancer behaves differently. There we can see that the difference in code length is larger for the subset of added occurrences than for the whole data. As we increase the time period, new occurrences worsen the overall model. As seen in Table 4, the increase in code length is largely due to the new occurrences.

For diabetes there are no big differences in code lengths between the time periods. Pairwise comparison in Table 4 shows that the improvement in code length is largest as we increase the time period from 180 days to 3 years. The comparison of three years to longer time periods indicates that we will not improve our model if we extend the time period. Again the differences between models were very small. The three year time period seems to be a reasonable choice also on basis of Table 4.

The worst code length for both of our mortality sequences is 28797, which would mean that we are not able to compress our original data at all. With the models used in this paper we obtain a code length which is approximately half of the worst code length. If we compress both of the mortalities with the Lempel-Ziv algorithm (Ziv and Lempel, 1978), we can get an idea of the size of the entropy for the sequences. With Lempel-Ziv the code length obtained for 90 days mortality is 3321 bits and 4219 bits for 365 days mortality. This means that both of the sequences could still be compressed much more than we were able to do with the logistic regression model. On the other hand, our compression with the logistic regression model uses information from the explanatory variables while Lempel-Ziv uses the sequence itself. Therefore comparison based solely on compression capability is not fair for the method presented in this paper.

6. Discussion

In this paper we have presented a sNML model selection criterion for logistic regression. The sequential approach enables us to compute the normalized maximum likelihood criterion also for large datasets. This was previously not possible for logistic regression models because of computational difficulties in the

normalizing constant of the NML criterion.

If the data doesn't have a natural ordering, we have to find one. This ordering should make sense from the applications perspective, which may be difficult in some cases. With the hip fracture data a natural (although not unique) ordering was obtained by using the date of arrival to hospital. The sequential approach also enables us to assume that the covariate effects develop over time. By using a sliding window in the calculation of the code length we are able to take this development in covariate effects into account and if necessary try to find the most suitable window length. If we approach the problem non-sequentially, we have to assume that covariate effects stood constant over the data.

Our objective was to find how far back in time we should look for three different comorbidities to get a good model for the mortality of hip fracture patients. We viewed each comorbidity separately from the other comorbidities. In our analysis we found out that we should use a different time period for each comorbidity. The results from sNML, AIC, BIC and c-statistic all pointed to the same direction for the choice of time period. This is a good result because the agreement of the different methods gives us stronger confirmation on the behavior of the comorbidities as explanatory variables. It also seems that in this case the sequential approach gives results which are in line to non-sequential approaches.

Our results indicate that for congestive heart failure we should use all medical history available to us, while for cancer it is enough to use only records from half a year before the fracture. For diabetes the message is not clear, but using records from three years before the fracture seems to be a reasonable choice. The results obtained by using a sliding window do not change our previous conclusions on the effect of different comorbidities. This suggests that there has not been any remarkable changes in covariate effects within the time period under consideration.

We were also able to distinguish how much of the change in codelength is due to the observations that become new indications of a comorbidity as we increase the time period that we look back in time. In congestive heart failure the fit of the whole data improves as we get new indications of that comorbidity. On the other hand, with cancer the model fits worse especially among the new cancer indications. Also this suggests that these two comorbidities behave in a quite different manner from each other.

All of the comorbidities improved the model. If we use the codelength obtained with Lempel-Ziv algorithm as a yardstick how far we are from the entropy, we can see that there is still a lot to improve. However, we do not want to lose interpretations about the explanatory variables effects on the outcome. Therefore we cannot construct a method aiming purely for maximum compression of data.

Acknowledgements

The work of Reijo Sund was financially supported by the Yrjö Jahnsson Foundation (grant, 5978).

Appendix

We give the algorithm for the computation of sNML (see (7)) in logistic regression. The mortality sequence $\{y_1, y_2, \dots, y_t\}$ is denoted as \mathbf{y}^t and the matrix $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_t\}$ of explanatory variables as X^t . Let n be the number of observations in the whole dataset and $y^{t|a}$ is the sequence of length t where the last observation $y_t = a$. First $m = r$ must be chosen in such a way that the sequentially maximized likelihood is finite (see e.g. Albert and Anderson, 1984). Then calculate the regular NML from the r observations. Denote this by $\hat{p}(\mathbf{y}^r|X^r)$. Compute the sequential part as:

0. Initialize $\Delta = -\log_2[\hat{p}(\mathbf{y}^r|X^r)]$
 1. For $i = (r + 1) : n$
 - 1.1 Solve the ML-estimate $\hat{\beta}_0$ by using $\mathbf{y}^{i|0}$ and X^i (see (3))
 - 1.2 Solve the ML-estimate $\hat{\beta}_1$ by using $\mathbf{y}^{i|1}$ and X^i
 - 1.3 Compute
$$P(y_t = 0|\mathbf{y}^{t-1}, X^t, \hat{\beta}_0(\mathbf{y}^t)) = 1/(1 + e^{\hat{\beta}_0^T \mathbf{x}_t}),$$

$$P(y_t = 1|\mathbf{y}^{t-1}, X^t, \hat{\beta}_1(\mathbf{y}^t)) = (e^{\hat{\beta}_1^T \mathbf{x}_t})/(1 + e^{\hat{\beta}_1^T \mathbf{x}_t})$$
 and
$$K(\mathbf{y}^{t-1}) = P(y_t = 0|\mathbf{y}^{t-1}, X^t, \hat{\beta}_0(\mathbf{y}^t)) + P(y_t = 1|\mathbf{y}^{t-1}, X^t, \hat{\beta}_1(\mathbf{y}^t))$$
 - 1.4 If $y_t = 0$ then
$$\Delta_i = -\log_2 P(y_t = 0|\mathbf{y}^{t-1}, X^t, \hat{\beta}_0(\mathbf{y}^t)) + \log_2 K(\mathbf{y}^{t-1})$$
else
$$\Delta_i = -\log_2 P(y_t = 1|\mathbf{y}^{t-1}, X^t, \hat{\beta}_1(\mathbf{y}^t)) + \log_2 K(\mathbf{y}^{t-1})$$
 - 1.5 Set $\Delta = \Delta + \Delta_i$

The codelength for the data is Δ .

References

- Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71**, 1-10.
- Barron, A., Rissanen J. and Yu, B. (1998). The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory* **44**, 2743-2760.

- Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, New York.
- DeLong, E. R., Peterson, E. D., DeLong, D. M., Muhlbaier, L. H., Hackett, S. and Mark, D. B. (1997). Comparing risk-adjustment methods for provider profiling. *Statistics in Medicine* **16**, 2645-2664.
- Hannan, E. L., Magaziner J., Wang, J. J., Eastwood, E. A., Silberzweig, S. B., Gilbert, M., Morrison, R. S., McLaughlin, M. A., Orosz, G. M. and Siu, A. L. (2001). Mortality and locomotion 6 months after hospitalization for hip fracture: risk factors and risk-adjusted hospital outcomes. *Journal of the American Medical Association* **285**, 2736-2742.
- Heithoff, H. A. and Lohr, K. N. (1990). Hip fracture: setting priorities for effective research. Report of a study by the Institute of Medicine, Division of Health Care Services, National Academy of Sciences, National Academy press, Washington, District of Columbia, 61-64.
- Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*, 2nd edition. Wiley, New York.
- Iezzoni, L. I. (1994). Using risk-adjusted outcomes to assess clinical practice: an overview of issues pertaining to risk adjustment. *Annals of Thoracic Surgery* **58**, 1822-1826.
- Iezzoni, L. I. (2003). Range of risk factors. In *Risk Adjustment for Measuring Health Care Outcomes* (Edited by L. I. Iezzoni), 3rd edition. Health Administration Press, Chicago.
- Keene, G. S., Parker, M. J. and Pryor, G. A. (1993). Mortality and morbidity after hip fractures. *British Medical Journal* **307**, 1248-1250.
- Landon, B., Iezzoni, L. I., Ash, A. S., Schwartz, M., Daley, J., Hughes, J. S. and Mackiernan, Y. D. (1996). Judging hospitals by severity-adjusted mortality rates: the case of CABG surgery. *Inquiry* **33**, 155-166.
- McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear and Mixed Models*. Wiley, New York.
- Normand, S-L. T., Glickman, M. E. and Gatsonis, C. A. (1997). Statistical methods for profiling providers of medical care: issues and applications. *Journal of the American Statistical Association* **92**, 803-814.

- Preen, D. B., Holman, C. D. J., Spilsbury, K., Semmens, J. B. and Brameld, K. J. (2006). Length of comorbidity lookback period affected regression model performance of administrative health data. *Journal of Clinical Epidemiology* **59**, 940-946.
- Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi, J. C., Saunders, L. D., Beck, C. A., Feasby, T. E. and Ghali, W. A. (2005). Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Medical Care* **43**, 1130-1139.
- Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory* **42**, 40-47.
- Rissanen, J. (2007). *Information and Complexity in Statistical Modeling*. Springer, New York.
- Rissanen J. and Roos T. (2007). Conditional NML universal models. In *Proceedings 2007 Information Theory and Applications Workshop, IEEE press*, 337-341.
- Roos T. and Rissanen J. (2008). On sequentially normalized maximum likelihood models. *Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-08)*. Tampere, Finland, August 18-20.
- Rosenbaum, P.D. (2002). *Observational Studies*. Springer, New York.
- Salem-Schatz, S., Moore, G., Rucker, M. and Pearson, S. D. (1994). The case for case-mix adjustment in practice profiling: when good apples look bad. *Journal of the American Medical Association* **272**, 871-874.
- Sund, R. (2007). Utilization of routinely collected administrative data in monitoring the incidence of aging dependent hip fracture. *Epidemiologic Perspectives & Innovations*, **4**:2.
<http://www.epi-perspectives.com/content/4/1/2>
- Sund, R., Nurmi-Lüthje, I., Lüthje, P., Tanninen, S., Narinen, A. and Keskimäki, I. (2007). Comparing properties of audit data and routinely collected register data in case of performance assessment of hip fracture treatment in Finland. *Methods of Information in Medicine* **46**, 558-566.
- Tabus, I. and Rissanen, J. (2006). Normalized maximum likelihood models for logit regression. In *Festschrift for Tarmo Pukkila on his 60th Birthday* (Edited by Liski, E. P., Isotalo, J., Niemelä, J., Puntanen, S., and Styan, G. P. H.), 159-172. University of Tampere, Department of Mathematics, Statistics and Philosophy, Report A 368.

Ziv, J. and Lempel, A. (1978). Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory* **24**, 530-536.

Received September 22, 2010; accepted January 12, 2011.

Antti Liski
Department of Signal Processing
Tampere University of Technology
PO Box 553, 33101 Tampere, Finland
antti.liski@tut.fi

Ioan Tăbuș
Department of Signal Processing
Tampere University of Technology
PO Box 553, 33101 Tampere, Finland
ioan.tabus@tut.fi

Reijo Sund
Service Systems Research Unit
National Institute for Health and Welfare
THL, PO Box 30, FI-00271 Helsinki, Finland
reijo.sund@thl.fi

Unto Häkkinen
Center for Health Economics
National Institute for Health and Welfare
THL, PO Box 30, FI-00271 Helsinki, Finland
unto.hakkinen@thl.fi

Tampereen teknillinen yliopisto
PL 527
33101 Tampere

Tampere University of Technology
P.O.B. 527
FI-33101 Tampere, Finland

ISBN 978-952-15-3041-8
ISSN 1459-2045